Centre for Research and Analysis of Migration

CReAM

Discussion Paper Series

CDP No 06/12

The role of language in shaping international migration

Alicia Adsera and Mariola Pytlikova

Centre for Research and Analysis of Migration
Department of Economics, University College London
Drayton House, 30 Gordon Street, London WC1H 0AX

CReAM Discussion Paper No 06/12

The role of language in shaping international migration

Alicia Adsera* and Mariola Pytlikova†

* Princeton University and IZA
† Aarhus University, CCP, CIM and CReAM

Non-Technical Abstract

Fluency in (or ease to quickly learn) the language of the destination country plays a key role in the transfer of human capital from the source country to another country and boosts the immigrant's rate of success at the destination's labor market. This suggests that the ability to learn and speak a foreign language might be an important factor in the migration decision. We use a novel dataset on immigration flows and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2009 and a wide range of linguistic indicators to study the role of language in shaping international migration. Specifically, we investigate how both linguistic distance and linguistic diversity, as a proxy for the "potential" ease to learn a new language and to adapt to a new context, affect migration. We find that migration rates increase with linguistic proximity and the result is robust to the inclusion of genetic distance as a proxy for cultural proximity and to the use of multiple measures of linguistic distance. Interestingly, linguistic proximity matters more for migrants moving into non-English speaking destinations than to English-speaking countries. The likely higher proficiency of the average migrant in English rather than in other languages may diminish the relevance of the linguistic proximity indicators to English speaking destinations. Finally, destinations that are linguistically more diverse and polarized attract fewer migrants than those with a single language; whereas more linguistic polarization at origin seems to act as a push factor.

Keywords: International migration, language.

JEL Classification: J61, F22, O15.

Centre for Research and Analysis of Migration
Department of Economics, Drayton House, 30 Gordon Street, London WC1H 0AX
Telephone Number: +44 (0)20 7679 5888
Facsimile Number: +44 (0)20 7679 1068

The role of language in shaping international migration

BY Alicia Adsera and Mariola Pytlikova*

January 2012

Abstract:

Fluency in (or ease to quickly learn) the language of the destination country plays a key role in the transfer of human capital from the source country to another country and boosts the immigrant's rate of success at the destination's labor market. This suggests that the ability to learn and speak a foreign language might be an important factor in the migration decision. We use a novel dataset on immigration flows and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2009 and a wide range of linguistic indicators to study the role of language in shaping international migration. Specifically, we investigate how both linguistic distance and linguistic diversity, as a proxy for the "potential" ease to learn a new language and to adapt to a new context, affect migration. We find that migration rates increase with linguistic proximity and the result is robust to the inclusion of genetic distance as a proxy for cultural proximity and to the use of multiple measures of linguistic distance. Interestingly, linguistic proximity matters more for migrants moving into non-English speaking destinations than to English-speaking countries. The likely higher proficiency of the average migrant in English rather than in other languages may diminish the relevance of the linguistic proximity indicators to English speaking destinations. Finally, destinations that are linguistically more diverse and polarized attract fewer migrants than those with a single language; whereas more linguistic polarization at origin seems to act as a push factor.

JEL Classification: J61, F22, O15

Keywords: International migration, language.

* Adsera: Princeton University and IZA; Woodrow Wilson School, 347 Wallace Hall, Princeton University, NJ 08544 (email: adsera@princeton.edu); Pytlikova: Aarhus University, CCP, CIM and CReAM, Department of Economics and Business, Frichshuset Hermodsvej 22, 8230 Åbyhøj (E-mail: marp@asb.dk). We are grateful to participants at NORFACE and CReAM conference on Migration: Economic Change, Social Challenge in London, EEA conference in Oslo, WISE International Symposium on Contemporary Labor Economics in Xiamen, IZA AM2 2010 in Bonn, First Tempo Conference on International Migration in Dublin and NORFACE, WB and CReAM conference on "Migration, Development, and Global Issues" in London, in particular to Bernt Bratsberg, Barry Chiswick, Christian Dustmann, Tor Eriksson, Gordon Hanson, Tim Hatton, Giovanni Peri, Chad Sparber, Jens Suedekum and Ott Toomet (alphabetical order), for their comments, and to Bo Honoré and Ingo E. Isphording, for helpful discussions. Jan Bryla, Enric Boix, Raymond Hicks, Anne Lund and Rasmus Steffensen provided excellent research assistance within the migration and linguistic data collection process. We also thank Ignacio Ortuno and Roman Wacziarg for providing the data on linguistic diversity and genetic distance. This research was funded in part by the NORFACE migration program. The usual disclaimer applies.

I. Introduction

Previous literature has shown that both fluency in the language of the destination country or the ability to learn it quickly play a key role in the transfer of existing human capital to a foreign country and generally boost the immigrant's success at the destination country's labor market, see e.g. Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007, 2010), Dustman (1994), Dustman and van Soest (2001 and 2002), and Dustman and Fabbri (2003). By exploiting differences between young and old arrivers from non-English speaking source countries on their adult English proficiency, Bleakley and Chin (2004 and 2010) find that linguistic competence is a key variable to explain immigrant's disparities in terms of educational attainment, earnings and social outcomes. Thus linguistic skills seem to be very important in accounting for migrants' well-being. Recent studies show that it is easier for a foreigner to acquire a language if her native language is linguistically closer to the language to be learned (Chiswick and Miller, 2005; Isphording and Otten, 2011). This suggests that the ability to learn and speak a foreign language quickly might be an important factor in the potential migrants' decision making. Besides, a "widely-spoken" native language in the destination country can constitute a pull-factor in international migration. Two different forces may lie behind that migration pattern. First, as some "widely spoken" languages are often taught as second languages at schools in many source countries, immigrants are more likely to move to destinations, where those languages are spoken in order to lower the costs associated with skill transferability and to increase their chances of being successful at the destination labour market. Second, foreign language proficiency may be considered an important part of human capital in the labor market of the source country, see e.g. European Commission (2002) on language proficiency as an essential skill for finding a job in home countries. A recent article by Toomet (2011) finds that knowledge of English is associated with a 15% wage premium on the Estonian labor market. Thus, learning/practising/improving the skills of "widely spoken" languages in the destination countries may serve as a pull factor especially for temporary migrants. Additionally the richness and variety of the linguistic environment where an individual is brought up may enhance his/her future ability to adapt to a new milieu. Numerous neuroscience and biology studies have argued that a multilingual environment may shape brains of children differently and increase their capacity to better absorb a larger number of languages (Kovacs and Mehler, 2009). If this is the case we should expect that, ceteris paribus, individuals from multi-lingual countries would have an easier time absorbing a new linguistic register in their destination country. In that regard the migration costs of those individuals would be

smaller than otherwise and we would expect larger immigration fluxes (and better outcomes, something beyond the scope of this paper) from those source countries, other things being constant. Regarding the effect of multi-lingual destinations on migration, there might be two forces pulling the effect into different directions: a linguistically polarized society may increase the costs of adaptation, but a diverse society might have in place more flexible policies that adapt to the needs of different constituencies (e.g. education, integration programs). Although the role of language and linguistic proximity seem to be very important, previous evidence on the determinants of migration typically included only a simple dummy for sharing a common language.[1]

The main contribution of this paper is to investigate in depth the role of language in shaping international migration by using a wide range of linguistic indicators and a novel international migration data. First, we examine the relevance of linguistic proximity between origin and destination countries in the decision to migrate and to this aim *we construct a set of refined indicators of the linguistic proximity* between two countries based on the linguistic family of either the first official, any other official or the major local language in each country. In addition, we investigate the role of linguistic proximity using two existing indices: first, the *Levenshtein linguistic distance* developed by the Max Planck Institute for Evolutionary Anthropology which relies on phonetic dissimilarity of words in two languages and, second, the *linguistic proximity measure proposed by Dyen et al.* (1992), a group of linguists who built an index of distance between Indo-European languages based on the similarity between samples of words from each language. To separate the relevance of language proximity on its own from other sources of cultural proximity we also include information on the genetic distance between the populations of the destination and the origin countries in the models. Second, we investigate the hypothesis that potential migrants prefer to choose a destination with a *"widely spoken" language*, such as English, as its local language. Third, we investigate the role of the richness and variety of the linguistic environment at destination and origin in the migration process. We also add to the existing literature on determinants of migration by using a rich novel international migration dataset, which allows us to analyze migration from a multi-country perspective. In this paper, we analyze the determinants of the annual gross migration flows from 223 countries to 30 OECD countries for the period 1980-2009.

[1] A few studies have also employed some more sophisticated linguistic measures. For instance Belot and Hatton (2012) use the number of nodes between one language and another on the linguistic tree to construct a linguistic proximity measure. Further, a recent paper by Belot and Ederveen (2012) employs the linguistic proximity index proposed by Dyen et al. (1992). The authors show that cultural barriers explain patterns of migration flows between developed countries better than traditional economic variables. In our paper, we use the Dyen index as a part of robustness analyses.

We find that emigration rates are higher among countries whose languages are more similar. The result is robust to the inclusion of genetic distance, which suggests language itself affects migration costs beyond any ease derived from moving to a destination where people may look or be culturally more similar to the migrant. We conduct the analysis by looking separately at both the proximity between the first official languages and between the major languages in each country as well as the maximum proximity between any of the official languages (if multiple) in both countries. We find that emigration flows to a country with the same language as opposed to those to a country with the most distant language are around 27% higher, ceteris paribus, and around 14% higher in the short-run in models that include the lagged dependent variable in addition to a large set of controls and time and country dummies. This result is highly robust to the use of alternative continuous measures of proximity developed by linguists both for the world sample (Levenshtein distance) and among countries with Indo-European languages (Dyen index). The implied increase in emigration rates to countries with similar language as opposite to linguistically distant countries ranges between 18.8 to 20%. When estimating separate coefficients on linguistic distance for English and non-English speaking destinations, linguistic proximity matters more for the latter group. The average migrant likely has some English proficiency, even before the move, that may temper the relevance of the linguistic proximity when studying flows to English speaking destinations. It might be that the return to English is higher in linguistically more distant countries, which in turn fuels temporary migration from those countries to English-speaking destinations. Finally, destinations that are linguistically more diverse and polarized attract fewer migrants than those with a single language; whereas more linguistic polarization at origin seems to act as a push factor.

The rest of the paper is organized as follows: Section 2 shortly presents a model on international migration on which we base our empirical analysis. Sections 3 and 4 describe the empirical model as well as the database on migration flows and stocks collected for this study, linguistic measures and other independent variables included in the analyses. Results from the econometric estimates are given in Section 5. Finally, Section 6 offers some concluding remarks.

II. A Model of International Migration

To introduce our empirical specification we present a model of migration across different destinations. This model follows the "human capital investment" theoretical framework (Sjastaad, 1962) and its recent application in Grogger and Hanson (2011)[2], which we simplify since we are

[2] Or similar application in Ortega and Peri (2009).

only interested in explaining aggregate migration flows and we do not distinguish among different skill levels as they do. We assume that an individual *k* decides whether to stay in his/her country of origin *i* or whether to migrate from country *i* to any potential destination *j*, where $j = 1, 2, .., J$.

A potential immigrant maximizing his/her utility chooses to locate in the country where his/her utility is the highest among all available destination choices. The utility that migrant *k*, currently living in *i*, attains by moving to *j* is given by:

$$U_{kij} = f(y_{kj} - c_{kij}) + \varepsilon_{kij} \qquad (1)$$

where *f* is a strictly increasing continuous function of the difference between income in destination *j*, y_{kj}, and the cost of migrating from the home country *i* to *j*, c_{kij}. A simple example is given by $U_{kij} = \alpha(y_{kj} - c_{kij})$, where the utility function is assumed to be a linear function with $\alpha > 0$. The utility that individual *k* obtains by staying in *i* naturally does not include moving costs. We can write the probability of individual *k* from country *i* choosing a country *j* among *J* possible destinations as:

$$\Pr(j_k / i_k) = \Pr\left[U_{ijk} = \max(U_{ki1}, U_{ki2}, ..., U_{kiJ})\right] \qquad (2)$$

Assuming that ε_{kij} follows an *i.i.d.* extreme value distribution, we can apply the results in McFadden (1974) to write the log odds of migrating to destination country *j* versus staying in the source country *i* as:

$$\ln \frac{M_{ij}}{P_i} = \ln m_{ij} = \alpha[y_j - y_i] - \alpha c_{ij} \qquad (3)$$

where M_{ij} are flows of individuals from *i* to *j*; P_i are the stayers; and m_{ij} is the emigration rate from *i* to *j*. The probability of migration depends on the difference between income related to staying at home country *i* or migrating abroad *j* adjusted for costs of migration, that include both pecuniary and other non-monetary utility differences between the two locations, c_{ij}.Costs of moving to foreign country may be three fold: direct out-of-pocket costs of migrating and psychological costs of leaving own country, family and friends, and costs associated with a loss of skills due to skill transferability.

The results in McFadden (1974) rely on the assumption that the relative probabilities of two alternative locations only depend on the characteristics of those two alternatives. The independence

of irrelevant alternatives (IIA) assumption can be considered implausible in some contexts. The empirical analysis of our paper includes only OECD destinations and we only need that the IIA holds for these countries (McFadden, 1974; Grogger and Hanson, 2011). In the result sections we comment on some additional tests we undertake to show such an assumption is plausible here.

For the case where the individual's utility is logarithmic, we can rewrite (1) as:

$$U_{kij} = (y_{kj} - c_{kij})^{\lambda} \exp(\varepsilon_{kij}) \tag{4}$$

As before we assume that ε_{kij} follows *i.i.d.* extreme value distribution and λ>0. Using the approximation that, $\ln(y_j - c_{ij}) \approx \ln y_j - (c_{ij} / y_j)$, the log odds of migrating to destination country *j* versus staying in the source country *i* are written as follows:

$$\ln \frac{M_{ij}}{P_i} = \ln m_{ij} \approx \lambda[\ln y_j - \ln y_i] - \lambda C_{ij} \tag{5}$$

Migration costs $C_{ij} = (c_{ij} / y_j)$ are now expressed as a proportion of destination income.

Suppose that income in a location, y_k, can be defined in line with Harris and Todaro (1970) as wage times the probability of finding a job, $y = we$, where e denotes employment rate[3], w real earnings. Then the migration rate in (5) can be expressed in terms of employment rates and wages:[4]

$$\ln \frac{M_{ij}}{P_i} = \ln m_{ij} \approx \lambda[\ln w_{kj} + \ln e_{kj} - \ln w_{ki} - \ln e_{ki}] - \lambda C_{ij} \tag{6}$$

III. Empirical Model Specification

We base our econometric analysis on the model presented in the previous section. The model assumes that emigration rates to one destination are driven by difference in wages, employment rates between origin and destination countries, and the costs of migration. Specifically, our econometric model has the following form:

[3] The employment rate can be expressed as one minus the unemployment rate, $y = w(1-u)$.

[4] Suppose that income in a location, y_i can be defined as average earnings from employment and benefits received otherwise, $y_i = w_i e_i + (1-e_i)\tau_i$, where τ are net transfers. Similarly as in equation (8), the migration rate is approximated by:

$$\ln \frac{M_{ij}}{P_i} \approx \lambda[\ln e_{kj} + \ln[w_{kj} + (\tau_{kj}(\frac{1}{e_{kj}} - 1))] - \ln e_{ki} - \ln[w_{ki} + (\tau_{ki}(\frac{1}{e_{ki}} - 1))] - \lambda C_{ij}$$

$$\ln m_{ijt} = \gamma_1 + \gamma_2 \ln(GDP_j)_{t-1} + \gamma_3 \ln(GDP_i)_{t-1} + \gamma_4 \ln(GDP_i)^2_{t-1} +$$
$$+\gamma_5 \ln u_{jt-1} + \gamma_6 \ln u_{it-1} + \gamma_7 \ln pse_{jt-1} + \gamma_8 \ln s_{ijt-1} + \gamma_9 L_{ij} + \gamma_{10} D_{ij} + \qquad (7)$$
$$+\gamma_{11} \ln p_{ijt-1} + \gamma_{12} FH_{it-1} + \delta_j + \delta_i + \theta_t + \varepsilon_{ijt}$$

where m_{ijt} denotes gross flows of migrants from country *i* to country *j* divided by the population of the country of origin *i* at time *t*, where *i=1,…,223; j=1,…,30* and *t=1,...,30.* Similarly as previous studies we proxy wages by GDP per capita and employment prospects in the sending and receiving countries by unemployment rates, u_{jt} and u_{it}. The effect of GDP per capita in the source country on migration flows may be nonlinear since poverty constrains the ability to cover costs of migration. It has been shown in previous studies, e.g. Chiquiar and Hanson (2005), Hatton and Williamson (2005), Clark et al. (2007), and Pedersen et al. (2008), that source country's GDP per capita has an inverted U-shape effect on migration.[5] Therefore, the level of GDP per capita in the source country also enters the model in a quadratic form, $\ln(GDP_i)^2_{t-1}$, as a means to account for the non-linearity effects pointed by the theory. In addition to the economic determinants, Borjas (1999) argues that generous social security payment structures may play a role in migrants' decision making. Potential emigrants must take into account the probability of being unemployed in the destination country and generous welfare benefits in the destination country constitute a substitute of earnings during the period devoted to searching for a job.[6] We include public social expenditure as percentage of GDP, $\ln pse_{jt-1}$, as a proxy for the "welfare magnet" among explanatory variables.

Migration costs are determined by different factors. Generally, the larger the physical distance between two countries the higher are the direct migration costs associated with transportation. However, changes and improvements in communication technologies, internet, continued globalization of the economy and declining costs of transportation lead to a decline in direct costs of migration over time. Second, we expect that the larger the language barrier, the higher are the migration costs for an individual associated with a lower chance to transfer her skills and knowledge into the destination's labour market. Further, migration "networks" (i.e. networks of family members, friends and people of the same origin that already live in a host country) play an important role in lowering the direct and psychological migration costs (Massey et al., 1993; Munshi, 2003). The "networks" can provide potential migrants with the necessary help and

[5] At income levels beyond dire poverty, migration increases, but after GDP reaches a certain level, migration may again decrease because the economic incentives to migrate to other countries decline. This may be related to the fact that due to the data limitations previous studies looked only at North-North or South-North migration and not South-South migration. It might be that individuals from poorer countries migrate close home.
[6] Theoretically one may incorporate the welfare benefits in case of unemployment into the model (6), see the footnote 4 for the application.

information and thus facilitate the move and the adaptation of new immigrants into the new environment. Thus, we expect that c_{ij} the migration costs associated with migration from country *i* to country *j* averaged over all individuals *k* are larger with physical, cultural and linguistic distance between countries, but fall with the existence of migration networks. They also depend on specific destination and origin factors (such as immigration laws in destinations or credit market constrains in origins). In our empirical specification we use the number of foreign population from country *i* living in country *j* per population of the source country *i*, s_{ijt}, to control for the network of migrants. The linguistic variables, central to the main hypotheses in this paper, are covered in matrix *L.* In line with our hypotheses presented above we add a measure of linguistic distance between countries and measures of linguistic diversity in destinations and origins. We use three different linguistic distance measures, specifically: (1) our own measure, which we constructed based on information from Ethnologue and which we call *Linguistic Proximity* index, (2) *Levenshtein distance* developed by the Max Planck Institute for Evolutionary Anthropology and (3) *Dyen linguistic proximity* measure proposed by Dyen et al. (1992). To account for the diversity of languages in both the country of origin and destination we use the *fractionalization* and *polarization* indices from Desmet et al. (2009) and Desmet et al. (2011). All these variables are described in detail in the data section below. To control for the effect of distance, matrix D_{ij} includes the following variables into our empirical specification: *Log Distance in Kilometres* between the capital areas in the sending and receiving countries; a dummy variable to proxy for cultural similarity denoted *Neighbour Country* which takes a value of 1 if the two countries are neighbours and 0 otherwise; and finally the dummy variable, *Colony,* with value 1 for countries ever in colonial relationship, and 0 otherwise. Past colonial ties might have some influence on the cultural distance between countries, increase the information available and general knowledge about the potential destination country in the source country and thus lower migration costs and encourage migration flows between these countries.

Political pressure in the source country may also influence migration. Therefore, we include a couple of indices from *Freedom House,* which aim to separately measure the degree of freedom in political rights and civil liberties in each country. Each variable takes on values from one to seven, with one representing the highest degree of freedom and seven the lowest. We expect violated political rights and civil liberties to increase migration outflows in a given country. On the other hand, political restrictions may also impede outmigration. The Freedom House variables are included in the matrix FH_i and come into the equation in logs. Finally we include a variable that

captures the relative population size in destination with respect to origin, $\ln p_{ijt-1}$, in order to control for demographic developments.

All variables used in the estimations, except dummy variables and the linguistic proximity indices, are expressed in logarithms. In order to account for what information was available to the potential migrant at the time the decision whether to move or not was made, the relative differences in economic development and employment between origin and destination countries are lagged by one period. More importantly, there might be a problem of reverse causality if migration flows impact both earnings and employment.[7] Lagging the economic explanatory variables and treating them as predetermined is one way to reduce the risks of reverse causality in the model.[8]

We first estimate the model in equation (7) by OLS without any country specific effects starting from parsimonious to full specifications. All specifications contain a set of year dummies, θ_t, in order to control for common idiosyncratic shocks over the time period[9] and robust standard errors clustered at each pair of destination and source country. Next, we estimate full models, which contain country of destination and country of origin fixed effects. In the context of international migration research, the question of whether to account for destination- and origin-country specific effects, δ_j and $\delta_{i,}$ separately or whether to include pair of countries specific effects, $\delta_{ij\,j,}$ comes up regularly. Destination and origin country fixed effects might capture unobserved characteristics of immigration policy practices in each destination country, credit market constrains in origins, as well as climate, openness towards foreigners or culture in each country, among other things. On the other hand, pair-wise fixed effects might capture (unobserved) traditions, historical and cultural ties between a particular pair of destination and origin countries, as well as bilateral immigration policy schemes between those countries. However, since the main focus of the paper is on the effect that linguistic and cultural proximity have on migration, and the pair-wise fixed effects would be collinear with the variables of interest, our preferred specification includes separate destination and origin country fixed effects with robust Hubert/White/sandwich standard errors clustered across pairs of countries.[10]

[7] There is another huge stream of literature that focuses on the effect of immigration on the labour market, see e.g. Borjas (2003) and Card (2005).

[8] With regard to the migrants' network, the variable is problematic too since the stock is just a function of previous stock plus migration flows minus out-migration. Therefore, we also lag the stock of migrants and assume that the lagged stock is predetermined with respect to the current migration flows.

[9] In separate specifications, we used a linear trend instead of year dummies. Results were essentially identical and are available from the authors upon request.

[10] Even though most previous studies on migration determinants have used linear models with log-transformed variable, a few have chosen count models to fit the nonnegative dependent variable (e.g. Belot and Ederveen (2012) used negative binomial; Simpson and Sparber (2010) used Tobit and Poisson count models). We obtained similar estimates of the model using nonlinear least squares where the level of migration flows is explained by the exponential of the linear combination of all log-transformed independent variables without imposing any restrictions between the mean and the variance as some count models require.

We add a one to each observation of immigration flows and foreign population stocks prior to constructing emigration and stock rates, so that once taking logs we do not discard the "zero" observations. Simpson and Sparber (2010) discuss the "zero problem" in migration data. In our data only around 4.5 % of observations have a value of zero.[11] In the model specifications, we partly control for the likely persistence of migration flows by including the lagged stock of foreigners, which in fact by construction consists of previous migration flows. In order to control fully for this persistence, and to separate pure "networks" effects from the persistence effects caused by the outcomes of previous periods, in some specifications we add the lagged dependent variable, which introduces additional dynamics into the model, and allows us to interpret results from the short-run perspective.[12] The dynamic model to be estimated has the following form:

$$\begin{aligned} \ln m_{ijt} &= \gamma_1 + \gamma_2 \ln(GDP_j)_{t-1} + \gamma_3 \ln(GDP_i)_{t-1} + \gamma_4 \ln(GDP_i)^2_{t-1} + \\ &\gamma_5 \ln u_{jt-1} + \gamma_6 \ln u_{it-1} + \gamma_7 \ln pse_{jt-1} + \gamma_8 \ln s_{ijt-1} + \gamma_9 L_{ij} + \gamma_{10} D_{ij} + \\ &+\gamma_{11} \ln p_{ijt-1} + \gamma_{12} FH_{it-1} + \gamma_{13} \ln m_{ijt-1} + \delta_j + \delta_i + \theta_t + \varepsilon_{ijt} \end{aligned} \qquad (8)$$

There is a substantial literature discussing the potential bias and inconsistency of estimators in fixed or random panel data models in a dynamic framework, as well as solutions to that, see e.g. Arellano-Bond (1991). However, as in our model we control for fixed effects separately at the level of destinations and origins, and the dynamics are introduced on the level of country pairs, we do not run into these problems. In our result part we comment on both models without and with lagged dependent variables, as shown in equations (7) and (8), respectively.

IV. Data

A. International migration data

The analysis is based on a novel dataset on immigration flows and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2009. The dataset has been collected by writing to selected national statistical offices for majority of the OECD countries to request detailed yearly information on immigration flows and foreign population stocks by source

[11] This percentage is much lower than either the 95% of zero values that Simpson and Sparber (2010) face or the usually reported in the trade literature when estimating gravity models.

[12] In the theoretical model the dynamics can be introduced similarly as in Hatton (1995) through the assumption that a potential migrant forms his/her expectations about the future utility gains on the basis of past experience and information, and that the formation of expectations follows a geometric series of values. This dynamic term allow us to control for persistence in the level of migration from different locations and to show short-run effects of different variables on migration.

country in their respective country.[13] For three countries, Korea, Mexico and Turkey (and partly Japan), we obtained the data from the OECD International Migration Database, see the Online Appendix Tables A1 and A2 for a detailed overview on sources of migration data. Our international migration dataset presents substantial progress over that used in past research and over the existing datasets such as data by Docquier and Marfouk (2006)[14]; United Nations[15], OECD and the World Bank. First, contrary to the mentioned datasets, our data covers both migration flows and foreign population stocks.[16] Second the data is much more comprehensive with respect to destinations, origins and time due to our own effort with data gathering from particular statistical offices. For an overview of comprehensiveness of observations of flows and stocks across all destination countries over time, see the Online Appendix Table A3 and Table A4, respectively. It is apparent that the data becomes more comprehensive over time and thus missing observations become less of a problem for more recent years. In our dataset, as in the other existing datasets, different countries use different definitions of an "immigrant" and draw their migration statistics from different sources.[17] In particular for foreign population stock, we preferably use the definition based on country of birth, see the Online Appendix Tables A1 and A2 for a detailed overview of definitions and sources for data on immigration flows and foreign population stock, respectively.

B. Language

Linguistic distance

First, we created a measure that captures the linguistic proximity between two languages based on information from the encyclopaedia of languages *Ethnologue* (Lewis, 2009). The *Linguistic Proximity index* ranges from 0 to 1 depending on how many levels of the linguistic family tree the languages of both the destination and the source country share. We constructed the index in the

[13] The original migration dataset by Pedersen, Pytlikova and Smith (2008) covered 22 OECD destination and 129 source countries over the period of years 1989-2000, see Pedersen, Pytlikova and Smith (2008) for a detailed description of the dataset. For the purpose of this paper we additionally collected data from eight other OECD countries as additional destinations – Czech and Slovak Republics, Hungary, Poland, Ireland, Turkey, South Korea and Mexico and extended the number of countries of origin to cover the entire world. Further, we prolonged the existing time period to include the years 1980-1989 and 2001-2009.

[14] The international migration dataset by Docquier and Marfouk (2006) contains estimates of emigration stocks and rates by educational attainment for 195 source countries in 2000 and 174 source countries in 1990.

[15] The United Nations Global Migration Database (UNGMD) contains data on the foreign population stock by source country, sex and age. The data comes from different sources such as population censuses, population registers, nationally representative surveys or other official statistical sources from 221 countries in the world. For the 195 countries that include information on the international migrant stock for at least two points in time, interpolation or extrapolation was used to estimate the international migrant stock on 1 July of the reference years, namely 1990, 1995, 2000, 2005 and 2010 (UN, 2008). Regarding flows of migrants, the UN data contains annual data on the inflow of migrants by country of origin for 29 countries, based on national data sources. The data series cover in a very unbalanced fashion the period 1980 to 2008.

[16] Migration flows is the inflow of immigrants to a destination from a given origin in a given year. The definition usually covers immigrants coming for a period of half year or longer. Foreign population stock is a number of foreigners from a given country of origin (usually we use definition of country of birth to determine origin of the migrants) living in a destination in a given year. The foreign population stock data are dated ultimo.

[17] Thus our data, although in much lesser degree than the datasets by Docquier and Marfouk(2006), OECD, United Nations and the World Bank, bears some problems related to different sources of migration data (censuses, registers or labour force surveys), different definitions of foreigner (country of birth and citizenship) and unbalanced nature of the data due to missing observations for some countries of destinations and origins.

following way. First we defined weights: the first equal to 0.1 if two languages are related at the most aggregated linguistic tree level, e.g. Indo-European versus Uralic (Finnish, Estonian, Hungarian); the second equal to 0.15 if two languages belong to the same second- linguistic tree level, e.g. Germanic versus Slavic languages; the third equal to 0.20 if two languages belong to the same third linguistic tree level, e.g. Germanic West vs. Germanic North languages; and the fourth equal to 0.25 if both languages belong to the same fourth level of linguistic tree family, e.g. Scandinavian West (Icelandic) vs. Scandinavian East (Danish, Norwegian and Swedish), German vs. English, or ItaloWest (Italian, French, Spanish, Catalan and Portuguese) vs. RomanceEast (Romanian). Then, we constructed the linguistic proximity index as a sum of those four weights, and we set the index equal to 0 if two languages did not belong to any common language family, and equal to 1 if the two countries had a common language. Thus the linguistic proximity index equals 0.1 if two languages are only related at the most aggregated linguistic tree level, e.g. Indo-European languages; 0.25 if two languages belong to the same first and second- linguistic tree level, e.g. Germanic languages; 0.45 if two languages share the same first up to third linguistic tree level, e.g. Germanic North languages; and 0.7 if both languages share all four levels of linguistic tree family, e.g. Scandinavian East (Danish, Norwegian and Swedish).

In addition to our own *Linguistic Proximity* index, we also employ two alternative continuous measures of proximity developed by linguists. The first one is the Levenshtein linguistic distance produced by the Max Planck Institute for Evolutionary Anthropology, which relies on phonetic dissimilarity of words in two languages. The continuous index increases with the distance between languages. Linguists choose a core set of the 40 more common words across languages describing everyday life and items; then, express them in a phonetic transcription called ASJP code and finally compute the number of steps needed to move from one word expressed in one language to that same word expressed in the other language. For a detailed description of the method, see Bakker et al., 2009).[18] In our country sample the index ranges from 0 (when the two languages are the same) to a maximum of 106.39 (for the distance between Laos and Korea). The second one is a linguistic proximity measure proposed by Dyen et al. (1992), a group of linguists who built a continuous index between zero and 1000 of the distance between Indo-European languages based on the similarity of samples of words from each language. The index increases with similarity between languages and it is equal to 1000 when the two languages are the same. With these measures we build a matrix that contains continuous metrics of proximity between any pair of languages from

[18] The Levenshtein index has already been used as a useful tool to measure the extent of difficulty in learning the local language among migrants to Germany (Isphording and Otten 2011).

our destinations-source pairs and provides a better adjusted and smoother indicator of proximity than the standard dummies for common language used in most the literature. Nonetheless, the sample size in specifications that employ the Dyen variable is severely reduced since only countries with Indo-European languages are included. To link the linguistic proximity (or distance) measures to country pairs we first use the main official language in the country. In order to account for the existence of multiple official languages in some countries, we also create two separate sets of linguistic proximity measures: one is set at the maximum proximity between two countries using any of those official languages and a second measures the proximity between the most widely used language in each country (which in some cases is not the first official language). We use those two additional proximity indices in our robustness analyses.

Linguistic diversity

To account for the diversity of languages in both the country of origin and destination we use fractionalization and polarization indices from Desmet et al. (2011).[19] The linguistic fractionalization index computes the probability that two individuals chosen at random will belong to different linguistic groups and the index is maximized when each individual belongs to a different group. [20] Linguistic polarization, in contrast, is maximized when there are two groups of equal size. [21] So if a country A consists of two linguistically different groups that are of the same size and country B has three linguistic groups of equal size, then country B is more diverse, but less polarized than A.[22] In addition we use three more measures from Desmet et al. (2009), GI fractionalization[23] and ER polarization indexes[24], which control for the distances between different linguistic groups in addition to their shares in the population, and PH peripheral heterogeneity

---

[19] Desmet et al. (2011) use linguistic trees, describing the genealogical relationship between the entire set of 6,912 world languages, to compute measures of fractionalization and polarization at different levels of linguistic aggregation. A complete discussion about the measures can be found in their paper.

[20] In particular, for $k(m)=1...K(m)$ groups of size Φk(m), where m = 1...M denotes the level of aggregation at which the group shares are considered, the linguistic fractionalization is calculated as:

$$ELF(m)=1-\sum_{k(m)=1}^{N(m)}[\Phi k(m)]^2$$

[21] We use the polarization measure from Desmet at al. (2011) which is calculated as:

$$Pol(m)=4\sum_{k(m)=1}^{N(m)}[\Phi k(m)]^2[1-\Phi k(m)]$$

[22] Even though Desmet et al. (2011) calculate these indices for 15 different levels of aggregation, in the paper for space reasons we only use their measures at the 4th level of aggregation of linguistic families available in the linguistic classification of Ethnologue (e.g. German vs. English). The implied diversity of the index changes somewhat as the level of linguistic aggregation varies. Desmet et al. (2011) state in their paper that "When measured using the ELF index, the average degree of diversity rises as the level of aggregation falls, as expected. When measured using a polarization index, diversity falls at high levels of aggregation, and plateaus as aggregation falls further." (p.10).

[23] The GI index was proposed by Greenberg (1956). It computes the population weighted total distances between all groups and can be interpreted as the expected distance between two randomly selected individuals. It is essentially a generalization of ELF, whereby distances between different groups are taken into account. Note that for this index the maximal diversity need not be attained when all groups are of the same size because it also depends on the linguistic distance between those groups.

[24] ER index is a special case of the family of polarization indices started by Esteban and Ray (1994) that controls for distances between linguistic groups.

index, which can be seen as an intermediate index between fractionalization and polarization as it takes into account the distance between the center and the peripheral groups, but not between the peripheral groups themselves. Desmet et al. (2009) define the distances by the number of potential linguistic branches that are shared between the languages of two groups. Finally, in order to account for the intensity of multilingualism we include the number of indigenous languages at the linguistic tree level in a country spoken by a minimum of 5% of the country's population. The measures on number of languages at different linguistic levels, spoken by different percentages of a country's population were graciously provided by Ignacio Ortuno-Ortin.

C. Other variables helping to explain migration

Besides the information on linguistic proximity and diversity, the dataset contains additional variables, which may help to explain the migration flows between countries as mentioned in the previous section. These variables were collected from various sources (e.g. OECD, the World Bank and others). Table 1 contains definitions, and sources of all variables used and their summary statistics.

V. RESULTS

A. Linguistic proximity

Columns 1 to 5 in Table 2 present pooled OLS estimates of different model specifications from the most parsimonious model that only includes the linguistic proximity index and a constant to a full specification (excluding unemployment rates)[25]. The estimated coefficient for our variable of interest, the index of linguistic proximity, is significant and positive across all specifications. Thus, other things being equal, emigration flows between two countries are larger the closer their languages are. In column (1) the index of linguistic proximity on its own explains approximately 11.6% of the variance in emigration rates (adj. R-squared). The coefficient of 3.4 implies the increase in emigration rates to a destination with the same language compared to one whose language has not a single linguistic level in common with that of the source country should be at least in the order of 340%. Unsurprisingly as additional controls are included in the model, the size of the coefficient shrinks notably in size. The model in column (2) contains, in addition to the

[25] The reason for showing the results without the unemployment variables is that the source country unemployment rates impose the largest restriction with respect to the number of missing observations. By excluding unemployment variables we have twice the number of observations as compared to models that include unemployment in the full specification in addition to all pull and push factors.

linguistic distance, economic variables and relative population of both countries as well as the physical distance between their capitals. The coefficient of the linguistic proximity index decreases from 3.4 to about 1.7, and continues to be highly significant. These additional socio-economic variables are clearly relevant in explaining the emigration flows since they account for close to 37% of the variance. In column (3) we add measures of political and civil freedom in origin, dummies for past colonial relationship between both countries as well as an indicator of whether they share common borders. Countries are expected to be more tightly related and migration is expected to be less costly when they share a colonial past or are geographically close. Moreover, some former colonies may have adopted the language of their colonial power which we argue facilitates population movements between them. The coefficient of linguistic proximity is only slightly affected by the inclusion of these measures and stands around to 1.35 in column (3). In addition to economic, colonial or geographic ties, part of the influx of new migrants into a country may be fuelled by a reduction in the moving cost to that particular destination driven by the existence of local networks and bidirectional information between both countries. Clearly, in column (4) the stock of immigrants for the same origin in the destination country is positively and significantly associated with current migration flows. The explanatory power (adjusted R-squared) of the model increases from 52% to 83% when adding the lagged stock of immigrants, which indicates a strong role of network effects in driving international migration or some sort of historical path dependence in the flows. The coefficient of the linguistic proximity drops sharply to 0.16 when including the lagged stock of immigrants in column (4). To control for recent flows of immigrants to the country as in equation (8) we add the lagged dependent variable in column (5). The short-run impact of the linguistic proximity is 0.083 and highly significant. That is, emigration flows to a country with the same language as opposed to a country with the most distant language should be around 8.3% higher, ceteris paribus.

Besides the variables considered in our full model in column (5), there are other unobservable factors that shape international migration flows and that are characteristic of particular countries. To account for the unobserved country-specific heterogeneity, we add destination and origin country fixed-effects to the model in columns (6-8). The coefficient of linguistic proximity in the fully specified model with lagged dependent variable that includes these fixed effects in column (8) is 0.142, and remains highly significant at 1%. It implies that emigration flows to a country with the same language as opposed to a country with the most distant language should be around 14.2% higher, ceteris paribus. Thus in the short-run the difference in emigration rates to France from either

Zambia with a linguistic index of 0.1 or Sao Tome with a linguistic index of 0.7 and Benin that has French as an official language and a linguistic index of 1 (0.9 and 0.3 units larger than Zambia's and Sao Tome's, respectively) will be in the order of either 12.8% higher than Zambia's or 4.3% higher those from Sao Tome, ceteris paribus. If the lagged dependent variable is omitted, the implied difference is not surprisingly much larger, around 27%.

In Table 3 we present results of our full model specification and include information on unemployment rates both at origin and destination countries. The number of observations decreases from approximately 47,000 to around 25,500 compared to models in Table 2 due to missing observations for source country unemployment rates. In the first three columns we show OLS estimates. In columns (4) to (6) we include destination and source country fixed effects to the model. When comparing the pooled OLS results with the panel models that include fixed effects for destination and source countries, the overall impression is that the sign and statistical significance of the estimated coefficients for the linguistic proximity index are quite robust across the different specifications. The coefficients for the index of linguistic proximity in the fixed-effects model in both column (8) in Table 2 and column (6) in Table 3, which include the exact same variables except for unemployment rates, are quite close, 0.142 and 0.188 respectively, despite the large reduction in the sample size. The difference is somewhat larger for the models that do not include lagged dependent variable, 0.273 and 0.436 respectively.[26]

Turning our attention to the other control variables included in the models, the coefficients of emigration rates from the previous year are always positive and highly significant indicating continuity in the direction of migration flows. Similarly to other studies such as Bauer et al. (2005), Clark et al. (2007), Pedersen et al. (2008), McKenzie and Rapoport (2010) and Beine et al. (2011) we find network effects to be an important determinant of subsequent migration. The stock of immigrants from the same origin at a given destination is positively associated with larger flows but the size of the estimated coefficient decreases substantively when the lag of the dependent variable is included.[27] Results of the models with lagged dependent variable in Tables 2 to 3 indicate that a 10% increase in the stock of migrants from a certain country is associated with an increase of around 1.8-1.7% in the emigration rate from this country in the short-run, ceteris paribus. Implied emigration rates to countries with high GDP per capita are substantial in all estimates in Tables 2 and 3. Coefficients in the full models with fixed effects and migration rates imply that a 10%

[26] As a way to test whether the IIA assumption holds for OECD destinations, models were re-estimated by excluding one destination at a time. Results are stable and available from authors.

[27] We have also estimated all regressions with t-2 lags in the migration stock variable. Only the coefficient of that variable was slightly lower and the rest remained unchanged. Results are available from the authors upon request.

increase in the GDP of the destination country is associated with an increase in emigration rates of slightly over 10%. The GDP per capita of the source country enters both linearly and in a quadratic form in all regressions. Estimated coefficients of the last columns of Table 2 and 3 imply that the relationship between GDP per capita of the origin country and emigration rates is nonlinear. Emigration rates remain pretty stable (or decrease somewhat) as GDP increases within the range of countries with very low levels of GDP per capita. As of a level corresponding to low-middle income countries emigration rates increase along with GDP per capita, thought this effect is quite moderate.[28] In fixed effects estimates, emigration rates are significantly higher from countries with relatively high unemployment rates, other things being the same. The finding for the unemployment rate at destination is ambiguous since it flips from being negative in column (5) to significantly positive in column (6) once the lagged dependent variable is included. The latter result, even if apparently surprising, may be explained by the relatively high unemployment rates experienced in many European countries during this period as compared to other periods and to other areas of the OECD coupled with their comparatively large welfare states. Nonetheless country fixed-effects and time dummies as well as the measure of public social expenditure should be already capturing some of those differences. The increased mobility of labor within EU countries during these last decades as barriers were dismantled may also be part of the explanation. In line with the theoretical framework proposed by Borjas (1999) and contrary to existent empirical evidence e.g. Zavodny (1997), Pedersen et al. (2008) and Wadensjö (2007), among others, we find that the coefficients to public social expenditure are positive and significant in models with fixed effects in Tables 2 and 3.[29] At any rate social expenditures would only be relevant for migrants as long as they are entitled to receive them but some of the OECD countries have a few universal benefits policies to which anybody is eligible regardless of nationality.[30] Population ratio enters positively and significantly in all models in Tables 2 and 3 except in the last column of each table in the most complete specification with fixed effects where it becomes insignificant. Distance is clearly significantly associated with weaker emigration flows in all specifications. Colonial past is significantly associated with stronger emigration flows in all fixed effects models. In column (6) of Table 3, having a past colonial tie increases the emigration rates to that destination by around 16%. Emigration rates from countries with more restrictive political rights are significantly larger in some

[28] This point of inflexion occurs at around levels of $2,000 in Table 2 and $4,000 in Table 3. This is related to the fact that the sample used in Table 3 contains relatively richer countries and more recent observations on average than that of Table 2 given that the unavailability of unemployment rates in source countries limits the sample importantly.
[29] However the public social expenditure measure is inversely related to emigration rates in the OLS estimates.
[30] This is something we plan to investigate further in a separate paper.

specifications. Lack of political liberties seems to act as a push factor but coefficients fail to attain significance in most specifications. Conversely, civil rights seem to be more relevant to explain migration patterns. In Table 2 and most columns in Table 3, controlling for political rights, emigration rates seem to be larger in countries with better civil rights. Some of these rights may be associated with lower barriers to out-migration and geographic mobility.

B. Robustness

To see how robust our results are to alternative measures, we substitute our linguistic proximity index with two continuous measures of linguistic distance between countries. First, we use the Levenshtein distance developed by the Max Planck Institute for Evolutionary Anthropology, which relies on phonetic dissimilarity of words in two languages and, second, we employ the linguistic proximity index proposed by Dyen et al. (1992) that measures the closeness between Indo-European languages based on the similarity between samples of words from each language. Given that the Dyen index covers only Indo-European languages, our number of observations is reduced significantly from around 25,500 to only close to 15,000 in the full model. Results of the full model specification with country fixed effects are presented in Table 4a. Regressions in the upper end of the table do not include the lagged dependent variable (columns 1 to 8) and those presented in the lower part of the table do (columns 9 to 16). Columns (1) and (9) include estimates using the Levenshtein index calculated for the main official language in each country, and columns (2) and (10) contain similar estimates with the Dyen index instead. The Leivenshtein index indicates distance as opposed to proximity between languages. As a result the significant negative estimate in all specifications indicates that emigration rates are larger to countries whose languages are closer as measured by Leveinshtein. As noted the index ranges from zero to a bit over 106 in our sample. The estimated coefficients imply that emigration rates to countries with similar languages as opposed to those with an index of around 100 (quite dissimilar) should be around 20% higher using the estimates of column (9). It is interesting to note that the size of the implied effect is remarkably similar to that found with our own original proximity index.

Similarly, the Dyen index displays a significant positive coefficient in all econometric specifications. The implied magnitude of the increase in emigration rates when comparing a country with the same language (and a Dyen index of 1000) and a country with a rather dissimilar language (the minimum of around 100 found in our sample of Indo-European languages) is around 18% using the estimates of column (16). It is very interesting to find such similar results using the Dyen

estimate to those obtained with the other indices. First, the sample is restricted to likely more homogeneous countries, since it excludes those source or destination countries with non-Indo-European languages. Second, the Dyen index (as well as the Leveinshtein index) allows for greater variance across country-pairs than our original index since it measures more continuously the proximity between languages than the other indicators in the paper. As shown the magnitude of the coefficient, 0.0002 in the fixed effects model, is non-negligible. For example, the difference in emigration rates to an English speaking country from Nepal (with a Dyen of 157 with respect to English) as compared to those from Zambia (with a score of 1000) should be around 17%. The difference between migrants from either Argentina (with an index of 240) or Austria (with an index of 578) with respect to someone from Zambia should be in order of 15% and 8.5% respectively.[31]

As part of the robustness analyses, we extend the set of linguistic measures to include an index that takes into account the existence of multiple official languages and we compute the index at the maximum proximity between two countries using any of those languages ("all"). The literature has shown that migrants from different linguistic backgrounds self-select to different areas within destination countries with multiple languages according to the most widely used language in each area. Chiswick and Miller (1995), one of the most prominent examples of this line of research, show how migrants to Canada self-select to the province whose language is closer to their own because that enhances their labor market returns. Finally, with the same methodology we construct an index of linguistic proximity using instead the language most extensively used in the country (the "major" language) even if in some countries it is not among the official ones. The coefficients of the linguistic proximity when using the two alternative criteria are significant and positive. In columns (3) and (6) they are of very similar size as that in column (5) of Table 3. Those that include the lagged dependent variable are slightly smaller than the estimated coefficient in column (6) of Table 3, which contains the exact same specifications with the basic index. The implied increase in emigration rates from a country with the same major language compared to those from one country with no linguistic relation to the destination are around 11.6%. The size of the increase is around 17% when employing the minimum distance between any of the official languages, ceteris paribus. Similarly, results in Table 4a are very stable for the Levenshtein and Dyen index when calculated

[31] In separate estimates we have used the Dyen index and attached a zero value for the pairs of countries in which one language belongs to the family of Indo-European languages and the other does not. The estimated coefficient on the index of 0.00015 is, not surprisingly slightly lower in value in full sample specification (with around 25,500 observations) than when the sample is restricted to only Indo-European countries, but it still remains highly significant and implies a difference in emigration rates of around 15% between countries with the same language and those that do not share any level of the linguistic tree. Conversely, in separate models not presented here, both the estimated coefficients of our index of linguistic proximity and their significance are slightly larger and closer to the Dyen estimate when we use a sample restricted to only countries with Indo-European languages instead of the whole sample. The estimated coefficient of the proximity index in the full specification with fixed effects and lagged dependent variable is 0.26 in that sample. Results are available upon request.

both for the distance between the major languages and for any of the official languages, though the size of the coefficient decreases somewhat in the latter case for the Dyen.

As an additional robustness analysis we run a set of regressions with dummies that indicate whether the two main official languages share the same linguistic family separately for each level of the linguistic tree and also a dummy that indicates whether the same language is spoken in the two countries in order to depict non-linearities of the linguistic proximity index (if any). The results of regressions with destination and origin fixed effects are presented in Table 4b, columns (1) to (5) without the lagged dependent variable and in (6) to (10) with the lagged dependent variable, respectively. We observe that dummies for all levels of the linguistic family tree - except for the most aggregated (Indo-European vs. Uralic) – display a significant positive coefficient that increases with the level of the tree, with the largest one corresponding to the one that denotes that the same languages are spoken in the two countries, and the second largest for the fourth level of linguistic tree family (e.g. Scandinavian West vs. Scandinavian East).

Finally, one possible critique of the linguistic proximity index can be that it captures cultural proximity between countries. In order to separate the effects of language and culture, we include a couple of measures of the genetic distance between populations of both countries in our regressions. These indices, provided to us by Roman Wacziarg, are based on the work by Cavalli-Sforza, Menozzi, and Piazza (1994) and have been already been employed in other contexts to study, for example, cross-country differences in development (Spolaore and Wacziarg 2009). A detailed explanation of how the indices were constructed can be found in these two publications. The first index ("dominant") measures, for each pair of countries, the distance between the ethnic groups with the largest shares of population in each country. As the genetic index increases the larger are the differences between two populations. It takes a zero if the distributions of alleles in both populations are identical. The second index ("weighted") takes into account within-country subpopulations that are genetically distant and calculates the distance between both countries by taking into account the difference between each pair of genetic groups and weighting them by their shares. The index provides the expected genetic distance between two randomly selected individuals, one from each country.

Results are presented in Table 4c for the full specification with fixed effects. Again regression results in the upper section of the table do not include the lagged dependent variable and those in the lower section of the table do. The first two columns of each section show the coefficients for both measures of genetic distance when no index of linguistic proximity is included in the

regression. All coefficients are either effectively zeros or surprisingly slightly positive in column (9) indicating that stronger migration flows when the genetic distance is larger.[32] The rest of the specifications of the table adds to the first columns either our linguistic proximity index, the Leveinshtein or the Dyen index. All estimates presented in Table 4c show that all of our linguistic proximity results are robust to the inclusion of both measures of genetic distance. Coefficients for the different linguistic indices are essentially the same as those in Tables 3 and 4a. This suggests that language on its own affects migration costs beyond any ease derived from moving to a destination where people may look or be culturally more similar to the migrant.

C. The Role of Widely Spoken Languages

Our linguistic proximity index does not take completely into account the importance of the use of some widely spoken Indo-European languages (particularly English) in the media (TV, music) internet, business or everyday life and the high frequency of English as a choice of second language in schools. Therefore in Table 5 the models include separate indicators of linguistic proximity for non-English and for English speaking destinations in order to examine the role of English as a widely spoken language. If there is some "proficiency" advantage from knowing English as a second language, we expect that the linguistic proximity between native languages should matter more for non-English speaking destinations than for the others. Results in Table 5 seem to confirm this hypothesis. All linguistic proximity indices are strong predictors of emigration rates toward non-English speaking destinations. The coefficients of both the linguistic proximity index (in columns 1 and 7) and the Levenshtein index (in columns 2 and 8) for English destinations are smaller, though still significant, sizable and positive, than those for non-English destinations. This gives support to the hypothesis that people may still migrate to destinations with a widely spoken language even if their mother languages are linguistically far from that language. First, even if they do not regularly speak it at home, many migrants may have previous knowledge of a widely spoken language taught at schools and used in the internet and movies, particularly English (see special Eurobarometer study on languages by European Commission (2006), and Pytlikova (2006)). Second, foreign language proficiency is an important part of human capital in the labor market of source countries (see e.g. European Commission (2002) on language proficiency as an essential skill for finding a job in home countries). Those returns to widely spoken language proficiency may

[32] However if we restrict the sample to the relatively more homogeneous countries included in the Dyen dataset (that share Indo-European languages) the coefficients of the genetic distance variables turn negative but continue to be insignificant except for the coefficient on the weighted genetic index that is significantly negative when the lagged dependent variable is not included. Thus it seems that for relatively closer countries genetics matter more to explain migration flows than when we look at the complete sample of the world.

be higher in countries which are linguistically more far away from the widely spoken language. Thus learning/practicing/improving the skills of "widely spoken" language in the "native" countries serve as a pull factor especially for temporary migrants who may take this skill back home. Interestingly in columns (3) and (9) when we use the Dyen index instead, we do not find this difference in coefficients. We speculate that this may be due to the more selective nature of this sample that only includes more homogenous countries with Indo-European languages. In columns (4) and (10) we drop the unemployment rates form the model which affords a much larger sample. In line with our hypotheses, the estimated coefficient of the linguistic proximity for English destinations is substantially smaller and only significant in the model that includes the lagged dependent variable. The finding is similar in columns (7) and (11) when we use the linguistic proximity of the major language in the country instead. Finally in columns (6) and (12) we use the proximity index for the closest pair among all the official languages of each country. The coefficient for English destinations is now larger than for non-English. We believe that this is likely related to the fact that English and other colonial languages are (if not first) likely second or third official languages in many countries where they are not necessarily neither majoritarian nor widely known by the whole population but they may be taught in schools.[33]

D. Linguistic Diversity and Polarization

Table 6 includes a set of measures of the linguistic fractionalization and polarization of sending and receiving countries as defined in section 5. Each one of the boxes corresponds to a different model that, in addition to the two coefficients presented in the table, also includes covariates for linguistic proximity, network, economic conditions, distance and year dummies. Each model is first estimated without including the lagged dependent variable and then including it. None of the models includes fixed effects because the available diversity and polarization indices are constant for each country over time. The upper part of the table shows coefficients for the diversity of languages both at destination and origin using the log of the measures of fractionalization and polarization of languages at the 4th level of the linguistic tree (lnELF and lnPOL) obtained from Desmet (2011). Estimated coefficients from both fractionalization and polarization indices are fairly similar, even though the mean value of fractionalization is slightly higher than that of fractionalization in destination and conversely at origin. Coefficients for the diversity of languages at destination are negative and highly significant in all specifications. Ceteris paribus, the higher the linguistic

[33] In additional models available upon request we have also included measures of the number of computers per capita in the country to calculate the access to information about countries, or to infer exposure to English or other languages though internet and media use. All results remain unchanged.

diversity at destination, the smaller the migration flows. The mechanism behind this finding is subject of speculation but it may be related to fear from migrants that adaptation will be costly when not only one but more languages need to be learnt, even though places with a tradition of linguistic diversity are potentially welcoming to people with a different linguistic background. Conversely, the flows of migrants from countries with high linguistic diversity are larger than from those with more homogeneous linguistic environments. Multilingualism might be viewed as an asset that facilitates language acquisition at destination and lowers migration costs.

The second row in Table 6 includes regressions with diversity indices, both at destination and at origin, which take into account the linguistic distance between each pair of languages. The fractionalization is represented by the GI index from Desmet (2009), which takes into account the actual distance of languages and not only the particular linguistic family to which they belong as the ELF indices do. The polarization is now measured by ER index (of the family of polarization measures started by Esteban and Ray (1994)), which takes into account not only the different number of languages and their share of speakers but also the linguistic distance between each pair of languages. Interestingly, once we control for linguistic distances the coefficients to fractionalization and polarization differ. In particular, the coefficients to the ER polarization index become much larger in absolute terms, while coefficients to the GI fractionalization index become slightly smaller, even though the means and ranges of both measures are relatively similar. This finding seems to support the hypothesis that people do not want to invest into two very different languages. A more deeply polarized linguistic environment at destination seems to deter migration flows, other things being the same. Conversely, more polarized societies seem to significantly push larger number of people in the search of a new life elsewhere. Interestingly, if we exclude our index of linguistic proximity in separate results not presented here, the coefficients for both fractionalization and polarization in destination (with and without distances) become more negative. This may indicate that the negative effect of linguistic diversity is tempered by taking into account the distance of the immigrant's language to the main official language of the destination. Individuals may be less reluctant to move to a linguistically diverse destination if their own language is relatively close to one (or the main) language at destination.

We also run regressions with PH peripheral diversity index studied by Desmet et al. (2009), which also account for distances but not among all linguistic groups as in the previous indexes, but between the center and the peripheral groups. Not surprisingly the coefficients to the PH index lie somewhat between the coefficients of GI fractionalization and ER polarization.

Finally, in the third row of Table 6, the total number of indigenous languages at the second level of the linguistic tree that are spoken by at least 5 % of the population at the country of destination are consistently negatively associated with inflows. Conversely, emigration rates are stronger the larger the number of languages spoken in a source country.

VI. Conclusions and Further Steps

Fluency in the language of the destination country plays an important role in the transfer of human capital of migrants to a foreign country and generally it reduces migration costs and increases the rate of success of immigrant at the destination country's labor market. Recent studies show that it is easier for a foreigner to acquire a language if her native language is linguistically closer to the language to be learned (Chiswick and Miller, 2005; Isphording and Otten, 2011). This suggests that speaking a language, which is linguistically close to that of the destination country, might be an important factor in the potential migrants' decision of where to locate. Previous research has already shown that sharing a language is associated with larger population movements across countries. In this paper we use a novel dataset on immigration flows and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2009 to study the role of language in shaping international migration in more detail. Specifically, we investigate how linguistic distance and linguistic diversity, as a proxy for the "potential" ease to learn a new language and to adapt to a new context, affect migration. In addition to the large collection effort with the international migration data, we construct our own linguistic proximity measure, which is based on information from the encyclopaedia of languages Ethnologue. We focus not only on the first official language but also in any other official languages and in the most widely spoken language in each country.

We find that emigration rates are higher among countries whose languages are more similar. The result also holds both for the analysis of the proximity between the most used language in each country as well as for the minimum distance between any of the official languages in both countries. Among countries with Indo-European languages this result is highly robust to the use of an alternative continuous distance measure developed by Dyen et al., a group of linguists. Similarly the result prevails when we use the Levenshtein index, a continue measure of distance developed by the Max Planck institute for the majority of world languages. Further, the effect of linguistic proximity is robust to the inclusion of genetic distance, which suggests that language itself affects migration costs beyond any ease derived from moving to a destination where people may look or be

culturally more similar to the migrant. When estimating separate coefficients for English and non-English speaking destinations, linguistic proximity matters more for the latter group. The likely higher proficiency of the average migrant in English rather than in other languages may diminish the relevance of the linguistic proximity indicators to English speaking destinations. Unfortunately, our indices are unable to capture the familiarity of migrants with languages (such as English) that may have been learnt in school or though media use.[34] Additionally, positive selection of migrants to some destinations could imply over the average knowledge of second languages among those migrants. However, individual data would be required to study this. Finally, we find that destinations that are more linguistically diverse and polarized attract fewer migrants; whereas more linguistic polarization at origin seems to act as a push factor.

This is, to our knowledge, the first paper that disentangles the relationship between migration rates and language from different perspectives: by studying the role of linguistic distance, the role of widely spoken language and the role of linguistic diversity. We further contribute to the literature by constructing a new measure of linguistic distance and by using information on migration for a large set of origin and destination countries that spans for three decades.

[34] Also since the extent of dubbing varies across the world, future constructing a good measure of the exposure of residents in each country to original movies or TV shows could prove a very interesting piece of future research.

REFERENCES

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specifications for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58 (2): 279 – 299.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman 2009. "Adding Typology to Lexicostatistics: A Combined Approach to Language Classification." *Linguistic Typology* 13(1): 169-181.

Bauer, Thomas, Gil Epstein, and Ira Gang. 2005. "Enclaves, Language, and the Location Choice of Migrants." *Journal of Population Economics* 18(4): 649-662.

Beine, Michel, Frederic Docquier, and Caglar Ozden. 2011. "Diasporas." *Journal of Development Economics* 95 (1): 30-41.

Belot, Michele, and Sjef Ederveen. 2012. "Cultural and Institutional Barriers in Migration between OECD Countries." *Journal of Population Economics*, forthcoming.

Belot, Michele, and Timothy J. Hatton. 2012. "Skill Selection and Immigration in OECD Countries." *Scandinavian Journal of Economics*, forthcoming.

Bleakley Hoyt, and Aimee Chin. 2004. "Language Skills and Earnings: Evidence from Childhood Immigrants." *Review of Economics and Statistics* 84 (2): 481-496.

Bleakley Hoyt, and Aimee Chin. 2010. "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants." *American Economic Journal: Applied Economics* 2(1): 165-192.

Borjas, George J. 1999. "Immigration and Welfare Magnets." *Journal of Labour Economics* 17 (4): 607-637.

Borjas, George J. 2003. "The Labor Demand Curve Is Downward Sloping: Reexamining The Impact of Immigration on The Labor Market." *The Quarterly Journal of Economics* 118(4): 1335-1374.

Card, David. 2005. "Is the New Immigration Really so Bad?" *Economic Journal* 115 (507): 300-323.

Cavalli-Sforza, Luigi L., Paolo Menozzi, and Alberto Piazza. 1994. *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press.

Chiquiar, Daniel, and Gordon Hanson (2005): "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the U.S.." *Journal of Political Economy* 113 (2): 239–281.

Chiswick, Barry R., and Paul W. Miller. 1995. "The Endogeneity between Language and Earnings: International Analyses." *Journal of Labor Economics* 13 (2): 246-288.

Chiswick, Barry R., and Paul W. Miller. 2002. "Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle." *Journal of Population Economics* 15(1): 31-57.

Chiswick, Barry R., and Paul W. Miller. 2005. "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages." *Journal of Multilingual and Multicultural Development* 26(1): 1-11.

Chiswick, Barry R., and Paul W. Miller. 2007. "Computer Usage, Destination Language Proficiency and the Earnings of Natives and Immigrants." *Review of the Economics of the Household* 5 (2): 129-157.

Chiswick, Barry R., and Paul W. Miller. 2010. "Occupational Language Requirements and the Value of English in the US Labor Market." *Journal of Population Economics* 23(1): 353–372.

Clark, Ximena, Timothy J. Hatton, and Jeffrey G. Williamson. 2007. "Explaining U.S. Immigration, 1971-1998." *The Review of Economics and Statistics* 89(2): 359-373.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. 2011. "The Political Economy of Ethnolinguistic Cleavages." *Journal of Development Economics* 97 (2012): 322-338.

Desmet, Klaus, Ignacio Ortuño-Ortín and Shlomo Weber. 2009. "Linguistic Diversity and Redistribution.," *Journal of the European Economic Association* 7 (6): 1291-1318.

Docquier, A. Marfouk. 2006. Dataset. In C. Ozden and M. Schiff (eds). *International Migration, Remittances and Development*, Palgrave Macmillan: New York.

Dustmann, Christian. (1994). "Speaking Fluency, Writing Fluency and Earnings of Migrants." *Journal of Population Economics* 7: 133–56.

Dustmann, Christian, and Arthur van Soest. 2001. "Language Fluency and Earnings: Estimation with Misclassified Language Indicators." *The Review of Economics and Statistics* 83 (4): 663-674.

Dustmann, Christian, and Arthur van Soest. 2002. "Language and the Earnings of Immigrants." *Industrial and Labor Relations Review* 55 (3):473–492.

Dustmann, Christian, and Francesca Fabbri.2003. "Language Proficiency and Labour Market Performance of Immigrants in the UK." Economic Journal113: 695-717.

Dyen Isidore, Kruskal Joseph B. and Paul Black. 1992. "*An Indo-European classification: A lexicostatistical experiment*." Transactions of the American Philosophical Society 82, Part5. Philadelphia.

Esteban, Joan Maria, and Debraj Ray. 1994. "On the Measurement of Polarization." *Econometrica* 62 (4): 819-851.

European Commission. 2002. "Candidate Countries Eurobarometer (CCB) http://europa.eu.int/comm/public_opinion/cceb_en.htm

European Commission. 2006. Special Eurobarometer on "Europeans and their languages"http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_sum_en.pdf

Greenberg, Joseph H. 1956. "The measurement of linguistic diversity." *Language* 32: 109-115.

Grogger, Jeffrey, and Gordon H. Hanson. 2011. "Income maximization and the selection and sorting of international migrants." *Journal of Development Economics* 95 (1): 42-57.

Hatton, Timothy J., and Jeffrey G. Williamson. 2005. "What Fundamentals Drive World Migration?" in George. Borjas and J. Crisp (eds), *Poverty, International Migration and Asylum*, Palgrave-Macmillan.

Harris, John R., and Michael P. Todaro.1970. "Migration, unemployment and development: A two-sector analysis." *American Economic Review* 60 (5): 126–142.

Isphording, Ingo, and Sebastian Otten. 2011. "Linguistic Distance and the Language Fluency of Immigrants." Ruhr Economic Papers No. 274.

Kossoudji, Sherrie A. 1988. "The Impact of English Language Ability on the Labor Market Opportunities of Asian and Hispanic Immigrant Men." *Journal of Labor Economics* 6 (3): 205-228.

Kovacs, Agnes M., and Mehler, Jacques 2009. "Flexible Learning of Multiple Speech Structures in Bilingual Infants." *Science* 325: 611-612.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Massey, Douglas S., Joaquin Arango, Graeme Hugo, Ali Kouaci, Adela Pellegrino, and Edward J. Taylor 1993. "Theories of International Migration: A Review and Appraisal."*Population and Development Review* 19 (3): 431–466.

McFadden, Daniel, 1974. "The measurement of urban travel demand." *Journal of Public Economics* 3, 303–328.

McKenzie, David, and Hillel Rapoport. 2010. "Self-Selection Patterns In Mexico-U.S. Migration: The Role of Migration Networks." *The Review of Economics and Statistics* 92 (4): 811-821.

Munshi, Kaivan. 2003. "Networks in the modern economy: Mexican migrants in the US labor market." *The Quarterly Journal of Economics* 118 (2): 549–599.

Ortega, Francesc, and Giovanni Peri. 2009. "The Causes and Effects of International Migrations: Evidence from OECD Countries 1980-2005." National Bureau of Economic Research Working Paper 14833.

Pedersen, Peder J., Mariola Pytlikova and Nina Smith. 2008. "Selection and Network Effects – Migration Flows into OECD Countries, 1990-2000." *European Economic Review* 52(7): 1160-1186.

Pytlikova, Mariola 2006. "Where did Central and Eastern European Emigrants Go and Why?" unpublished manuscript.

Toomet, Ott. 2011. "Learn English, Not the Local Language! Ethnic Russians in the Baltic States." *American Economic Review* 101(3): 526–31.

Simpson, Nicole B. and Chad Sparber. 2010. "The Short- and Long-Run Determinants of Unskilled Immigration into U.S. States." Colgate University Working Paper 2010-06.

Spolaore, Enrico and Romain Wacziarg. 2009. "The Diffusion of Development." *Quarterly Journal of Economics* 124 (2): 469-530.

Sjastaad Larry A.. 1962. "The Costs and Returns of Human Migration." *Journal of Political Economy* 70 (5): 80-93.

United Nations, Department of Economic and Social Affairs, Population Division. 2008. "United Nations Global Migration Database (UNGMD)".

Wadensjö, Eskil. 2007. "Migration to Sweden from the New EU Member States." IZA Discussion Paper No. 3190.

Zavodny, Madeleine. 1997. "Welfare and the Locational Choices of New Immigrants." *Economic Review – Federal Reserve Bank of Dallas;* 2Q: 2-10.

Table 1: Descriptive statistics, definitions and sources

VARIABLES	Definition	Source	Obs	Mean	Sd	Min	Max
Linguistic Proximity	Linguistic Proximity index between i and j countries using their main official language.	Own calculation based on Ethnologue, see Data section	240840	.14002	.25005	0	1
Linguistic Proximity All	Linguistic Proximity index set at the maximum proximity between two countries using any of their official languages	Own calculation based on Ethnologue, see Data section	240840	.25467	.32527	0	1
Linguistic Proximity Major	Linguistic Proximity index between i and j countries using language spoken by majority	Own calculation based on Ethnologue, see Data section	240840	.08299	.19210	0	1
Dyen	Dyen Linguistic Proximity between i and j countries using their main official language based on the similarity of samples of words from each language	Dyen et al. (1992)	113184	414.3834	277.7418	110.6	1000
Dyen All	Dyen Linguistic Proximity set at the maximum proximity between two countries using any of their official languages	Dyen et al. (1992)	165672	490.5674	299.5441	112.8	1000
Dyen Major	Dyen Linguistic Proximity between i and j countries using language spoken by majority	Dyen et al. (1992)	75348	371.2698	260.85	110.6	1000
Levenshtein	Levenshtein linguistic distance between i and j countries using their main official language	Max Planck Institute for Evolutionary Anthropology	234360	87.6383	23.5913	0	106.39
Levenshtein All	Levenshtein linguistic distance set at the maximum proximity between two countries using any of their official languages	Max Planck Institute for Evolutionary Anthropology	238680	78.2922	30.3529	0	106.39
Levenshtein Major	Levenshtein linguistic distance between i and j countries using language spoken by majority	Max Planck Institute for Evolutionary Anthropology	219240	91.6020	18.6014	0	106.4
Ln Emigration Rate	Ln(migration inflow from i to j per source population)	Own data collection, see Tables A1 and A3	95408	-5.1221	2.5552	-14.0408	4.1193
Ln Emigration Rate_t-1	Ln(migration inflow from i to j per source population)t-1	Own data collection, see Tables A1 and A3	95408	-5.1220	2.5552	-14.0408	4.1193
Ln Stock of Migrants_t-1	Ln(foreign population stock from i in j per source population) t-1	Own data collection, , see Tables A2 and A4	75284	-3.1922	2.8966	-12.1770	6.5313
Ln Destination GDPperCap_t-1	Ln GDP per capita, PPP (constant 2005 international $) in destination j, t-1	WDI, World Bank	195348	10.0130	.4372	8.6204	11.2175
Ln Origin GDPperCap_t-1	Ln GDP per capita, PPP (const 2005 international $) in origin i, t-1	WDI, World Bank	146880	8.4735	1.2607	5.01600	11.4662
Ln Origin GDPperCap_t-1 sq	Ln GDP per capita, PPP (const 2005 intern $) in origin i squared, t-1	WDI, World Bank	146880	73.3894	21.3679	25.1603	131.4736
Ln Public Expenditure	Ln Public social expenditure as a percentage of GDP in destination j, t-1	OECD SOCX Database	165466	2.8618	.4920	.5038	3.5748
Ln Destination UnemplRate_t-1	Ln Unemployment, total (% of total labor force) in destination j, t-1	WDI, World Bank	172379	1.8382	.5534	.3924	3.1732
Ln Origin UnemplRate_t-1	Ln Unemployment, total (% of total labor force) in origin i, t-1	WDI, World Bank	73560	1.9767	.7149	-1.8707	4.0860
Ln Population Ratio, t-1	Ln Share of population in destination j per population in country i, t-1	WDI, World Bank	217590	8.3776	2.7334	-1.5113	17.3123
Ln Distance in km	Ln Distance between capitals of destination j and origin i in km	Own extension of CEPII	239724	8.5867	.8919	2.2741	9.8839
Neighboring Dummy	Dummy variable for neighbouring countries	Own extension of CEPII	240840	.01839	.1343	0	1
Historical Past Dummy	Dummy variable for countries ever in colonial relationship	Own extension of Rose (2004)	240840	.01779	.1322	0	1
Ln Origin Political Rights_t-1	Ln of Freedom House Index – Political Rights in origin i	Freedom in the World Scores	181740	1.0931	.7438	0	1.9459
Ln Origin Civil Rights_t-1	Ln of Freedom House Index – Civil Liberties in origin i	Freedom in the World Scores	181740	1.1501	.6450	0	1.9459
lnElfj4	Ln ELF fractionalization index in destination j	Desmet at al. (2011)	240840	-2.1792	1.1479	-5.5215	-.5459
lnPolj4	Ln Polarization index in destination j	Desmet at al. (2011)	240840	-1.5792	1.1122	-4.9619	-.07904
lnGIj	Ln GI fractionalization in destination j	Desmet at al. (2009)	232812	-3.0688	1.1809	-6.2146	-1.3509
lnERj	Ln ER polarization index in destination j, controls for the distances between different linguistic groups	Desmet at al. (2009)	232812	-3.8817	.9942	-6.2146	-2.2164
lnPHj	Ln PH peripheral heterogeneity index in destination j	Desmet at al. (2009)	232812	-3.1634	1.1423	-6.2146	-1.4024
lnELFi4	Ln ELF fractionalization index in origin i	Desmet at al. (2011)	223560	-1.3827	1.4788	-6.9078	-.0090
lnPoli4	Ln Polarization index in origin i	Desmet at al. (2011)	224640	-1.6025	1.7701	-6.9078	-.0032
lnGIi	Ln GI fractionalization in origin i	Desmet at al. (2009)	223560	-2.4583	1.5437	-6.9078	-.4293
lnERi	Ln ER polarization index in origin i, the index controls for the distances between different linguistic groups	Desmet at al. (2009)	223560	-3.6344	1.3076	-6.9078	-1.3863
lnPHi	Ln PH peripheral heterogeneity index in origin i	Desmet at al. (2009)	223560	-2.7735	1.4427	-6.9078	-.6912
LnN.LangMin5%j	No of indigenous languages at the 2nd linguistic tree level in *j* spoken by a minimum of 5% of population	Ignacio Ortuno-Ortin	240840	.2311	.3268	0	.6932
LnN.LangMin5%i	No of indigenous languages at the 2nd linguistic tree level in *i* spoken by a minimum of 5% of population	Ignacio Ortuno-Ortin	224640	.4422	.4616	0	1.7918
Dominant Genetic Distance	Dominant genetic distance between plurality groups, current match	Spolaore and Waciarg (2009)	233280	933.9762	720.1979	0	2760
Weighted Genetic Distance	Weighted genetic distance, current match	Spolaore and Waciarg (2009)	207576	941.7764	651.6594	0	2777.695

Table 2. Language proximity and migration rates from 223 countries of origin to 30 OECD destination countries for 1980-2009.

	OLS	OLS	OLS	OLS	OLS	FE	FE	FE
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Linguistic Proximity	3.362***	1.687***	1.355***	0.161**	0.083***	0.909***	0.273***	0.142***
	(0.151)	(0.119)	(0.121)	(0.066)	(0.020)	(0.124)	(0.069)	(0.026)
Ln Emigration Rate_t-1					0.744***			0.676***
					(0.009)			(0.011)
Ln Stock of Migrants_t-1				0.669***	0.158***		0.656***	0.179***
				(0.007)	(0.007)		(0.010)	(0.009)
Ln Destination		2.279***	2.305***	0.593***	0.186***	1.402***	2.422***	1.146***
GDPperCapPPPj_t-1		(0.075)	(0.074)	(0.049)	(0.017)	(0.158)	(0.148)	(0.082)
Ln Origin		1.185***	1.556***	0.627***	0.252***	1.445***	-0.103	-0.254**
GDPperCapPPPi_t-1		(0.261)	(0.257)	(0.143)	(0.043)	(0.332)	(0.317)	(0.120)
Ln Origin		-0.045***	-0.070***	-0.036***	-0.014***	-0.111***	0.001	0.018**
GDPperCapPPPit-1 squared		(0.016)	(0.015)	(0.009)	(0.003)	(0.021)	(0.020)	(0.008)
Ln Destination		-0.876***	-0.891***	-0.398***	-0.097***	0.662***	0.749***	0.375***
Public Social Expenditure_t-1		(0.094)	(0.091)	(0.062)	(0.017)	(0.100)	(0.097)	(0.053)
Ln Population Ratio_t-1		0.479***	0.476***	0.166***	0.045***	1.382***	0.460***	0.000
		(0.011)	(0.011)	(0.007)	(0.003)	(0.145)	(0.135)	(0.058)
Ln Distance in km		-0.642***	-0.605***	-0.258***	-0.097***	-1.067***	-0.408***	-0.173***
		(0.032)	(0.034)	(0.019)	(0.006)	(0.050)	(0.031)	(0.013)
Neighbouring Dummy			1.002***	0.054	-0.021	0.345**	-0.124	-0.054
			(0.163)	(0.097)	(0.029)	(0.161)	(0.091)	(0.033)
Historical Past Dummy			2.318***	-0.081	0.047	2.725***	0.511***	0.246***
			(0.218)	(0.164)	(0.046)	(0.193)	(0.149)	(0.052)
Ln Origin Freedom			-0.127**	0.038	0.009	0.016	0.023	0.015
Political Rightsi_t-1			(0.056)	(0.032)	(0.011)	(0.031)	(0.027)	(0.013)
Ln Origin Freedom			-0.100	-0.131***	-0.033**	-0.130***	-0.111***	-0.061***
Civil Rightsi_t-1			(0.070)	(0.042)	(0.014)	(0.039)	(0.032)	(0.017)
Destination & Origin FE	NO	NO	NO	NO	NO	YES	YES	YES
Constant	-5.733***	-30.852***	-32.440***	-9.694***	-3.078***	-25.215***	-29.731***	-11.727***
	(0.044)	(1.239)	(1.232)	(0.787)	(0.267)	(2.650)	(2.493)	(1.202)
Observations	95,408	74,805	72,100	47,910	46,004	72100	47910	46004
Adjusted R-squared	0.116	0.485	0.518	0.828	0.919	0.724	0.863	0.923

Notes: OLS estimates with and without fixed effects. Dependent Variable: Ln (Emigration Rate). All models include year dummies. Robust standard errors clustered at the country-pair level in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 3. Language proximity and migration rates from 223 countries of origin to 30 OECD destination countries for 1980-2009 with controls for unemployment rates.

	OLS	OLS	OLS	FE	FE	FE
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
Ln Linguistic Proximity	1.364***	0.279***	0.107***	1.205***	0.436***	0.188***
	(0.156)	(0.081)	(0.023)	(0.150)	(0.081)	(0.029)
Ln Emigration Rate_t-1			0.781***			0.707***
			(0.012)			(0.014)
Ln Stock of Migrants_t-1		0.719***	0.144***		0.679***	0.173***
		(0.009)	(0.009)		(0.012)	(0.011)
Ln Destination GDPperCapPPPj_t-1	3.077***	0.288***	0.149***	0.965***	2.130***	1.007***
	(0.113)	(0.075)	(0.023)	(0.213)	(0.189)	(0.103)
Ln Origin GDPperCapPPPi_t-1	1.067**	0.756***	0.288***	1.710***	-0.944*	-0.618***
	(0.421)	(0.238)	(0.072)	(0.582)	(0.488)	(0.199)
Ln Origin GDPperCapPPPit-1 squared	-0.041*	-0.045***	-0.017***	-0.091***	0.054*	0.037***
	(0.025)	(0.014)	(0.004)	(0.034)	(0.028)	(0.011)
Ln Destination Public Social Expenditure_t-1	-1.427***	-0.268***	-0.103***	0.541***	0.498***	0.252***
	(0.126)	(0.086)	(0.021)	(0.145)	(0.123)	(0.063)
Ln Destination UnemplRate_t-1	0.779***	-0.116***	0.040***	-0.033	-0.074**	0.060***
	(0.064)	(0.036)	(0.011)	(0.040)	(0.034)	(0.014)
Ln Origin UnemplRate_t-1	0.013	-0.053**	-0.016**	0.112***	0.082***	0.036**
	(0.046)	(0.026)	(0.008)	(0.029)	(0.026)	(0.014)
Ln Population Ratio_t-1	0.483***	0.117***	0.026***	1.668***	0.527***	0.036
	(0.014)	(0.009)	(0.003)	(0.193)	(0.169)	(0.075)
Ln Distance in km	-0.558***	-0.236***	-0.082***	-0.983***	-0.368***	-0.134***
	(0.037)	(0.022)	(0.007)	(0.049)	(0.035)	(0.013)
Neighbouring Dummy	0.923***	-0.066	-0.061**	0.308*	-0.229**	-0.084***
	(0.175)	(0.101)	(0.024)	(0.158)	(0.092)	(0.029)
Historical Past Dummy	2.201***	-0.236	0.025	2.567***	0.393*	0.161**
	(0.270)	(0.248)	(0.066)	(0.253)	(0.235)	(0.076)
Ln Origin Freedom PoliticalRi_t-1	0.048	0.039	-0.006	0.115***	0.077**	0.025
	(0.068)	(0.041)	(0.013)	(0.039)	(0.034)	(0.019)
Ln Origin Freedom CivilRi_t-1	-0.035	-0.130**	-0.025	-0.093**	-0.067*	-0.029
	(0.083)	(0.051)	(0.016)	(0.044)	(0.038)	(0.021)
Destination and Origin FE	NO	NO	NO	YES	YES	YES
Constant	-38.574***	-6.784***	-2.655***	-27.763***	-23.912***	-8.998***
	(2.005)	(1.281)	(0.392)	(4.013)	(3.454)	(1.647)
Observations	36165	26235	25408	36165	26235	25408
Adjusted R-squared	0.537	0.837	0.933	0.751	0.876	0.936

Notes: OLS estimates with and without fixed effects Dependent Variable: Ln (Emigration Rate). All models include year dummies. Robust standard errors clustered at the country-pair level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4a: Robustness checks: Alternative measures of linguistic proximity (Dyen and Levenshtein linguistic indexes and/or controls for multiple official languages) and migration rates to OECD countries.

	First Official Language		All Official Languages			Major Language		
Ling. Proximity/Distance measured by	(1) Levenshtein	(2) Dyen	(3) Ling.Proximity	(4) Levenshtein	(5) Dyen	(6) Ling.Proximity	(7) Levenshtein	(8) Dyen
Linguistic Proximity/Distance	-0.004***	0.0004***	0.368***	-0.004***	0.001***	0.481***	-0.004***	0.0005***
	(0.001)	(0.000)	(0.071)	(0.001)	(0.000)	(0.089)	(0.001)	(0.000)
Ln Emigration Rate_t-1	NO	NO	NO	NO	NO	NO	NO	NO
Constant	-23.761***	-11.300*	-24.341***	-23.531***	-12.13***	-6.713***	-24.089***	-11.501*
	(3.466)	(5.989)	(3.452)	(3.440)	(4.506)	(1.280)	(3.478)	(6.674)
Observations	25,770	15,301	26,235	26,180	19,970	26,235	25,841	13,170
Adj. R2	0.875	0.872	0.876	0.877	0.877	0.837	0.875	0.872
	First Official Language		All Official Languages			Major Language		
Ling. Proximity/Distance measured by	(9) Levenshtein	(10) Dyen	(11) Ling.Proximity	(12) Levenshtein	(13) Dyen	(14) Ling.Proximity	(15) Levenshtein	(16) Dyen
Linguistic Proximity/Distance	-0.002***	0.0002***	0.170***	-0.002***	0.0002***	0.116***	-0.002***	0.0002***
	(0.000)	(0.000)	(0.025)	(0.000)	(0.000)	(0.025)	(0.000)	(0.000)
Ln Emigration Rate_t-1	0.706***	0.705***	0.707***	0.707***	0.702***	0.781***	0.707***	0.704***
	(0.014)	(0.019)	(0.014)	(0.014)	(0.017)	(0.012)	(0.014)	(0.021)
Constant	-9.108***	-6.317**	-9.221***	-8.915***	-6.059***	-2.629***	-9.102***	-5.933*
	(1.656)	(2.891)	(1.643)	(1.639)	(2.186)	(0.391)	(1.663)	(3.239)
Observations	24,962	14,889	25,408	25,356	19,440	25,408	25,033	12,794
Adj. R2	0.935	0.932	0.936	0.936	0.935	0.933	0.936	0.932

Notes: Dependent Variable: Ln(Emigration Rate). Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 4b: Robustness checks: Linguistic families of first official language and migration rates to OECD countries.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Common Level 1	-0.032	-	-	-	-	-0.008	-	-	-	-
	(0.069)	-	-	-	-	(0.023)	-	-	-	-
Common Level 2	-	0.125***	-	-	-	-	0.059***	-	-	-
	-	(0.045)	-	-	-	-	(0.016)	-	-	-
Common Level 3	-	-	0.228***	-	-	-	-	0.096***	-	-
	-	-	(0.047)	-	-	-	-	(0.017)	-	-
Common Level 4	-	-	-	0.345***	-	-	-	-	0.138***	-
	-	-	-	(0.060)	-	-	-	-	(0.021)	-
Common Language	-	-	-	-	0.381***	-	-	-	-	0.167***
	-	-	-	-	(0.091)	-	-	-	-	(0.032)
Ln Emigration Rate_t-1	NO	NO	NO	NO	NO	0.710***	0.709***	0.708***	0.707***	0.708***
						(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Constant	-23.524***	-23.522***	-23.854***	-24.040***	-23.751***	-8.760***	-8.774***	-8.954***	-9.044***	-8.908***
	(3.435)	(3.444)	(3.453)	(3.448)	(3.437)	(1.639)	(1.642)	(1.645)	(1.648)	(1.641)
Observations	26,235	26,235	26,235	26,235	26,235	25,408	25,408	25,408	25,408	25,408
Adjusted R-squared	0.876	0.876	0.876	0.877	0.876	0.936	0.936	0.936	0.936	0.936

Notes: Dependent Variable: Ln (Emigration Rate). Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 4c: Robustness checks: Genetic Distance, Linguistic Proximity and Migration Rates to OECD countries

Linguistic Proximity/Distance measured by	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	No Linguistic Variable		Linguistic Proximity		Levenshtein		Dyen	
Linguistic Proximity/Distance			0.462***	0.458***	-0.004***	-0.004***	0.0004***	0.0004***
			(0.082)	(0.082)	(0.001)	(0.001)	(0.000)	(0.000)
Dominant Genetic Distance	0.000		0.000**		0.000*		-0.000	
	(0.000)		(0.000)		(0.000)		(0.000)	
Weighted Genetic Distance		0.000		0.000		0.000		-0.0004
		(0.000)		(0.000)		(0.000)		(0.000)
Ln Emigration Rate_t-1	NO	NO	NO	NO	NO	NO	NO	NO
Constant	-23.744***	-23.680***	-24.153***	-24.082***	-23.975***	-23.892***	-11.274*	-11.146*
	(3.431)	(3.427)	(3.450)	(3.448)	(3.462)	(3.466)	(5.997)	(5.979)
Observations	26,136	26,014	26136	26014	25,671	25,579	15,224	15,212
Adj. R2	0.876	0.876	0.877	0.877	0.875	0.875	0.872	0.872
Linguistic Proximity measured by	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	No Linguistic Variable		Linguistic Proximity		Levenshtein		Dyen	
Linguistic Proximity/Distance			0.197***	0.194***	-0.002***	-0.002***	0.0002***	0.0002***
			(0.029)	(0.029)	(0.000)	(0.000)	(0.000)	(0.000)
Dominant Genetic Distance	0.00004**		0.000***		0.000***		0.000	
	(0.000)		(0.000)		(0.000)		(0.000)	
Weighted Genetic Distance		0.000		0.000*		0.000		-0.000
		(0.000)		(0.000)		(0.000)		(0.000)
Ln Emigration Rate_t-1	0.709***	0.710***	0.707***	0.707***	0.706***	0.706***	0.706***	0.706***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.019)	(0.016)
Constant	-8.862***	-8.928***	-9.123***	-9.184***	-9.226***	-9.252***	-6.364**	-6.317**
	(1.639)	(1.638)	(1.649)	(1.649)	(1.658)	(1.657)	(2.893)	(2.893)
Observations	25,313	25,194	25313	25194	24,867	24,778	14,816	14,804
Adj. R2	0.936	0.936	0.936	0.937	0.936	0.936	0.932	0.932

Notes: Dependent Variable: Ln (Emigration Rate) from 223 countries of origin to 30 OECD destinations for 1980-2009.Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 5. The role of English as widely spoken language and migration rates to OECD countries.

	First Official Language				Major Language	All Official Languages
Linguistic Proximity/Distance measured by	(1) Ling.Proximity	(2) Levenshtein	(3) Dyen	(4) Ling.Proximity	(5) Ling.Proximity	(6) Ling.Proximity
Linguistic Proximity/Distance						
ln Non-English destination	0.538*** (0.082)	-0.005*** (0.001)	0.0003*** (0.000)	0.409*** (0.077)	0.620*** (0.086)	0.294*** (0.077)
ln English destination	0.283** (0.141)	-0.003** (0.001)	0.0008*** (0.000)	0.126 (0.106)	0.219 (0.175)	0.479*** (0.112)
Ln Emigration Rate_t-1	NO	NO	NO	NO	NO	NO
Other controls	YES	YES	YES	No Unemployment rates	YES	YES
Constant	-23.916*** (3.447)	-23.818*** (3.469)	-11.404* (5.969)	-29.734*** (2.499)	-23.949*** (3.449)	-24.276*** (3.464)
Observations *Adj. R2*	26,235 0.877	25,770 0.875	15,301 0.872	47,910 0.863	26,235 0.877	26,235 0.876
	First Official Language				Major Language	All Official Languages
Linguistic Proximity/Distance measured by	(7) Ling.Proximity	(8) Levenshtein	(9) Dyen	(10) Ling.Proximity	(11) Ling.Proximity	(12) Ling.Proximity
Linguistic Proximity/Distance						
ln Non-English destination	0.230*** (0.031)	-0.002*** (0.000)	0.0002*** (0.000)	0.208*** (0.030)	0.250*** (0.032)	0.147*** (0.027)
ln English destination	0.126*** (0.046)	-0.001*** (0.000)	0.0003*** (0.000)	0.073* (0.038)	0.068 (0.56)	0.205*** (0.037)
Ln Emigration Rate_t-1	0.707*** (0.014)	0.706*** (0.014)	0.704*** (0.019)	0.675*** (0.011)	0.706*** (0.014)	0.707*** (0.014)
Other controls	YES	YES	YES	No Unemployment rates	YES	YES
Constant	-9.000*** (1.647)	-9.133*** (1.657)	-6.381** (2.891)	-11.733*** (1.204)	-8.995*** (1.64)	-9.216*** (1.643)
Observations *Adj. R2*	25,408 0.936	24,962 0.935	14,889 0.932	46,004 0.923	25,408 0.937	25,408 0.936

Notes: Dependent Variable: Ln (Emigration Rate). Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 6. Linguistic diversity and polarization in destinations and origins and migration rates to OECD countries.

	(1)	(2)	(3)	(4)
Linguistic diversity (LD) measured by: In:	LnELF – a diversity index without distances		LnPOL- a polarization index without distances	
LD in Destination	-0.020 (0.018)	-0.013***(0.005)	-0.020 (0.019)	-0.013***(0.005)
LD in Origin	0.040***(0.012)	0.013***(0.003)	0.039***(0.013)	0.014***(0.004)
Ln Emigration Rate_t-1	NO	0.779***(0.012)	NO	0.779***(0.012)
Observations	26211	25386	26211	25386
Adj. R2	0.838	0.934	0.838	0.934
LD measured by:	LnGI - a diversity index with distances		LnER - a polarization index with distances	
LD in Destination	-0.019 (0.016)	-0.012**(0.005)	-0.050***(0.018)	-0.021***(0.005)
LD in Origin	0.022*(0.013)	0.006 (0.004)	0.028*(0.015)	0.009**(0.004)
Ln Emigration Rate_t-1	NO	0.773***(0.012)	NO	0.772***(0.012)
Observations	24204	23391	24204	23391
Adj. R2	0.841	0.934	0.841	0.934
LD measured by:	LnPH– peripheral diversity index		LnN.LangMin5%- N. languages at tree level 2 spoken by at least 5% population	
LD in Destination	-0.026 (0.016)	-0.013***(0.005)	-0.296***(0.055)	-0.115***(0.015)
LD in Origin	0.023 (0.014)	0.007*(0.004)	0.071**(0.036)	0.020*(0.010)
Ln Emigration Rate_t-1	NO	0.773***(0.012)	NO	0.777***(0.012)
Observations	24204	23391	26211	25386
Adj. R2	0.841	0.934	0.839	0.934

Note: Dependent Variable: Ln(Emigration Rate). Controls included: linguistic proximity index, stock of migrants, economic variables, distance variables and year dummies. Robust standard errors clustered at the country-pair level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix

Appendix Table A1: Inflows of Foreign Population: Definitions and Sources

Migration flows to:	*Definition of "foreigner" based on*	*Source*
Australia	Country of Birth	Permanent and long term arrivals, Government of Australia, DIMA, Dept. of Immigration and Multicultural Affairs http://www.immi.gov.au/media/statistics/index.htm
Austria	Citizenship	Population register, Statistik Austria (1997 to 2002), Wanderungsstatistik 1996-2001, Vienna
Belgium	Citizenship	Population register. Institut National de Statistique.
Canada	Country of Birth	Issues of permanent residence permit. Statistics Canada – Citizenship and Immigration Statistics. *Flow is defined as a sum of foreign students, foreign workers and permanent residents.* http://www.cic.gc.ca/english/resources/statistics/facts2009/glossary.asp
Czech Rep.	Citizenship	Permanent residence permit and long-term visa, Population register, Czech Statistical Office
Denmark	Citizenship	Population register. Danmarks Statistics
Finland	Citizenship	Population register. Finish central statistical office
France	Citizenship	Statistics on long-term migration produced by the 'Institut national d'études démographiques (INED)' on the base on residence permit data (validity at least 1 year) transmitted by the Ministry of Interior.
Germany	Citizenship	Population register. Statistisches Bundesamt
Greece	Citizenship	Labour force survey. National Statistical Service of Greece 2006-2007 Eurostat
Hungary	Citizenship	Residence permits, National Hungary statistical office.
Iceland	Citizenship	Population register. Hagstofa Islands national statistical office.
Ireland	Country of Birth	Labour Force Survey. Central Statistical Office. Very aggregate, only very few individual origins.
Italy	Citizenship	Residence Permits. ISTAT
Japan	Citizenship	Years 1988-2005: Permanent and long-term permits. Register of Foreigners, Ministry of Justice, Office of Immigration. Years 2006-2008: Permanent and long-term permits. OECD Source International Migration data
Korea	Citizenship	OECD Source International Migration data
Luxembourg	Citizenship	Population register, Statistical Office Luxembourg
Mexico	Citizenship	OECD Source International Migration data
Netherlands	Country of Birth	Population register, CBS
New Zealand	Last Permanent Residence	Permanent and Long-term ARRIVALS (Annual – Dec) Census, Statistics New Zealand
Norway	1979-1984 Country of Origin 1985-2009 Citizenship	Population register, Statistics Norway
Poland	Country of Origin	Administrative systems (PESEL, POBYT), statistical surveys (LFS, EU-SILC, Population censuses). Central Statistical Office of Poland
Portugal	Citizenship	Residence Permit, Ministry of Interior.
Slovak rep.	Country of Origin	Permanent residence permit and long-term visa, Slovak Statistical Office
Spain	Country of Origin	Residence Permit, Ministry of Interior
Sweden	Citizenship	Population register, Statistics Sweden
Switzerland	Citizenship	Register of Foreigners, Federal Foreign Office of Switzerland
Turkey	Citizenship	OECD Source International Migration data
United Kingdom	Citizenship	Residence permits for at least 12 months. IPS - office for national statistics, and EUROSTAT
United States	Country of Birth	US Census Bureau Current Population Survey (CPS); U.S. Department of Homeland Security: *Yearbook of Immigration Statistics*. Persons obtaining Legal Permanent Resident Status by Region and Country of birth www.dhs.gov/ximgtn/statistics/publications/LPR06.shtm)

Appendix Table A2: Stock of Foreign Population: Definitions and Sources

Foreign population stock in:	*Definition of "foreigner" based on*	*Source*
Australia	Country of birth	Census of Population and Housing, Australian Bureau of Statistics
Austria	Country of birth	Statistics Austria, Population Census 2001 and Population Register 2001 to 2009. For census year 1981 and 1991 definition by citizenship
Belgium	Citizenship	Population register. Institut National de Statistique
Canada	Country of birth	Census of Canada, Statistics Canada. www.statcan.ca/
Czech Rep.	Citizenship	Permanent residence permit and long-term visa, Population register, Czech Statistical Office and Directorate of Alien and Border Police
Denmark	Country of origin	Population register. Danmarks Statistics
Finland	Country of birth	Population register. Finish central statistical office
France	Country of birth	Census. Residence permit. Office des migrations internationals.
Germany	Citizenship	Population register. Statistisches Bundesamt
Greece	Citizenship	Labour force survey. National Statistical Service of Greece.
Hungary	Citizenship	National Hungary statistical office
Iceland	Country of birth	Population register. Hagstofa Islands
Ireland	Country of birth	Censuses, Statistical office, Ireland
Italy	Citizenship	Residence Permits. ISTAT
Japan	Citizenship	Years 1980-1999, Register of Foreigners, Ministry of Justice, Office of Immigration. Years 1999-2008 OECD Source Migration stat. Both sources based on permanent and long-term permits.
Korea	Citizenship	1986-1988: Trends in international migration Outlook, OECD 1990-2008: OECD Source International Migration Database
Luxembourg	Citizenship	Population register, Statistical office Luxembourg
Mexico	Country of birth	2005: Trends in international migration Outlook, OECD 2000: OECD Source International Migration Database
Netherlands	Citizenship	Population register, CBS
New Zealand	Country of birth	Census, Statistics New Zealand
Norway	Country background	Population register, Statistics Norway Country background is the person's own, their mother's or possibly their father's country of birth. Persons without an immigrant background only have Norway (000) as their country background. In cases where the parents have different countries of birth, the mother's country of birth is chosen.
Poland	Country of birth	2002 Census, rest permits, Statistics Poland
Portugal	Citizenship	Residence Permit, Ministry of Interior, www.ine.pt
Slovak Republic	Country of Origin	Permanent residence permit and long-term visa, Slovak Statistical Office
Spain	1985-1995 Citizenship 1996-2009 Country of birth	Residence Permit, Ministry of Interior
Sweden	Country of Birth	Population register, Statistics Sweden
Switzerland	Citizenship	Register of Foreigners, Federal Foreign Office
Turkey	Country of birth	OECD Source International Migration Database
United Kingdom	Country of Birth	LFS, UK statistical office
United States	Country of birth	US Census Bureau: 1990 and 2000 US census, the rest Current Population Survey (CPS) December. Data Ferret. Years 1980-1989, 1991-2004 from extrapolations by Tim Hatton (RESTAT)

Appendix Table A3: Country-Year Coverage migration flows

Columns: Destination Countries

Rows: Year

Cell: numbers of source countries, for which we have observations on the number of migrants for particular year

Dest	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	MEX	NLD	NOR	NZL	POL	PRT	SVK	SE	TUR	USA
Year																														
2009	205	190		218	195	141	193	203	113	183				139	2	179				141		198	213	212	123	150	212	194		194
2008	204	190		218	195	143	196	203	113	183	120			142	2	179	187	218	57	146	126	195	213	213	205	143	212	194	200	194
2007	206	190	93	218	195	147	195	203	113	183	124		192	128	2	179	181	218	28	142	126	197	213	213	205	126	211	194	199	196
2006	206	190	96	218	195	142	195	203	108	183	120	19	191	133	2	179	182	199	10	139		193	213	213	205	128	208	194	199	193
2005	203	190	85	218	195	142	195	203	66	183	107	178		121	2	179	185	10	10	137		187	213	213	205	124	208	194	199	195
2004	203	190	71	218	195	146	195	203	57	183	107	176		108	2	179	183	10	10	135		193	213	213	205	118	208	194	199	206
2003	201	198	70	218	195	142	195	203	57	183	127	176		122	2	179	180	10	10	127		191	213	213	205	114	208	194	199	206
2002	198	198	70	218	195	141	195	203	57	183	128	175		111	2	179	182	10	10	123		198	192	213	205	126	208	194	199	206
2001	198	198	70	218	195	115	195	203	57	183	130	195		117	2	179	181	10	10	116		197	192	213	205	114	208	194	200	206
2000	200	198	70	218	180	110	195	203	59	183	129	127		118	2	179	182	15	10	124		197	192	213	205	113	208	194	200	206
1999	198	198	70	218	179	108	195	203	58	183	118	127		114	2	179	181	15		123		191	192	213	205	114	208	160	200	206
1998	193	198	70	218	179	122	195	203	59	183	117	131	189	114	2	179	182	14		120		191	192	213	16	144	208	166	200	206
1997	192	198	55	218	179	111	195	203	39	183	118	9	184	114	2	179	179	14		110		194	192	213	14	144	208	164	200	206
1996	195	198	55	218	175	114	195	203	58	183	118	10	206	116	2	179	178	14		108		191	191	213	14	144	208	167	200	206
1995	187		55	218	175	117	195	203	39	183	118	7	204	117	2	179	48	15		110		187	192	213	13	144		165	200	206
1994	186		55	218	178	106	195	203	39	183	118	5	206	120	2	179	32	14		103		186	192	213	13	144		164		206
1993	180		48	218	177	97	195	203	39	183		6	206	107	2	179	32	14		99		185	192	213	11	143		168		206
1992	182		48	218	173		195	203	45	183		9	206	112	2	179	32	14		105		174	191	213	11	143		157		206
1991	171		48	218	157		195	203	42	183		7	206	105	2	179	32	11		95		160	191	213	11			148		206
1990	168		48	218	156		195	203	42	183		38	201	103	2	179	32	12		100		163	190	213	10			144		206
1989	155		48	218	154		195	203	42	183		31		98	2	179	32	11		93		164	192	213	10			142		206
1988	150		25	218	159		195	203	42	183		38		101	2	179	32	11		94		158	192	213				138		206
1987	159		27	218	155		195	203		183		29		99	2	179	32			93		161	192	213				136		206
1986	153		27	218	154		194	203		183		33		104		179	32						191	213				138		206
1985	155		27	218	154		195	203		183		35		95		18	32						116	213				134		206
1984	154		27	218	151		194	203		183						18							214	213				126		206
1983	166		27	218	152		195	203		183						18							214	213				123		206
1982	161		27	218	154		195	203								18							214	213				121		206
1981			27	218	154		195	203								18							214	213				123		206
1980			27	218			195	203															214	213				119		204

Appendix Table A4: Country-Year Coverage migration stocks

Columns: Destination Countries

Rows: Year

Cell: numbers of source countries, for which we have observations on the number of migrants for particular year

Dest	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	MEX	NLD	NOR	NZL	POL	PRT	SVK	SE	TUR	USA
Year																														
2009	209	209	185		195	172	190	201	113	191		171		180		175	190	12		26		207	213		209	177	145	200		133
2008	209	209	187		195	171	192	201	113	191		177		178		175	192	202	28	26		209	213			176	144	200		133
2007	209	209	178		195	168	193	200	113	191	128	174		174		175	188	201	25	26		207	213			179	142	200		133
2006	200	209	184	210	195	168	193	200	113	193	193	148	190	173	43	175	189	199	25	23		207	213	211		174	144	200		96
2005	209	209	182		195	166	139	201	113	193		97	192	165		175	189	183	25	23	10	208	213			173	139	200		96
2004	208	209	182		195	165	139	201	113	193		101	190	162		172	188	18	25	23		208	213			171	137	200		96
2003	208	209	182		195	163	139	201	113	193		100	191	156		172	188	18	25	23		207	213			168	149	200		96
2002	208	209	182		195	161	139	201	99	193		100		158	177	172	186	42	25	23		207	213		201	168	148	200		96
2001	190	207	182	190	195	163	139	201	99	193		97		154		172	187	42	19	12		206	213	199		167	142	200		96
2000	207	191	177		195	161	139	201	99	193		102	209	163		172	184	122	19	137	202	206	213			164	140	200	22	133
1999	206		175		195	164	139	201	99	193	162	87		163		172	185	42	19	12		204	213			158	136	111		96
1998	206		175		195	158	139	201	99	193		104		161		172	38	42	19	12		204	213			155	144	111		96
1997	204		55		195	152	139	201	99	193		100	190	159		172	189	42	19	12		204	213			152	144	111		96
1996	192		55	201	195	153	139	201	63	193		90	206	157	36	65	50	18	19	12		204	213	52		151	139	111		96
1995	202		55		195	150	139	201	58	193		85	206	146		65	50	37	19	12		200	213			151	140	111		96
1994	49		55		195	145	139	201	58	193		87	206			66	50	18	19	12		9	213			147		107		126
1993	49		48		195		139	201	58	193		87	206			66	50	18	19	12		9	213			140		104		126
1992	49		48		195		139	201	58	193		82	206			66	191	18	17	12		9	213			130		101		126
1991	168		48	180	195		136	201	58	193		70	206		2	43	190	16	15	12		9	213	51		126		98		126
1990	49	70	48		195		118	201	57	193	76		206			60		42	15	82		9	213			121		100	12	127
1989			48		195		118	201	57	134			206			60		12		8		9	213			122		98		125
1988					195		118	201	57	134			206			60		12	3	8		9	213			120		98		125
1987					195		118	201	57	131			206			60		12	5	8		9	213			118		97		125
1986	75			42	195		118	201	57	125			206		2	60		12	9	8		9	213	75		115		94		125
1985					195		118	201	57	124			206			60		42				9	213			109		95		125
1984					195		117	201		191			206			60		12				9	195			103		89		125
1983					195		118	201					206			60		12				9	195			100				125
1982					195		118	201					206			60		12					202			83		85		125
1981	81		47	42	195		118	201					206		2	59		12					198	75		98				125
1980		64			195		116	201					206					42		79			199			90		95		128